Spring 5-31-2012

Registration and categorization of camera captured documents

Venkata Gopal Edupuganti
New Jersey Institute of Technology

Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

REGISTRATION AND CATEGORIZATION OF CAMERA CAPTURED DOCUMENTS

by
Venkata Gopal Edupuganti

Camera captured document image analysis concerns with processing of documents captured with hand-held sensors, smart phones, or other capturing devices using advanced image processing, computer vision, pattern recognition, and machine learning techniques. As there is no constrained capturing in the real world, the captured documents suffer from illumination variation, viewpoint variation, highly variable scale/resolution, background clutter, occlusion, and non-rigid deformations e.g., folds and crumples. Document registration is a problem where the image of a template document whose layout is known is registered with a test document image. Literature in camera captured document mosaicing addressed the registration of captured documents with the assumption of considerable amount of single chunk overlapping content. These methods cannot be directly applied to registration of forms, bills, and other commercial documents where the fixed content is distributed into tiny portions across the document. On the other hand, most of the existing document image registration methods work with scanned documents under affine transformation. Literature in document image retrieval addressed categorization of documents based on text, figures, etc. However, the scalability of existing document categorization methodologies based on logo identification is very limited. This dissertation focuses on two problems (i) registration of captured documents where the overlapping content is distributed into tiny portions across the documents and (ii) categorization of captured documents into predefined logo classes that scale to large datasets using local invariant features.

A novel methodology is proposed for the registration of user defined Regions Of Interest (ROI) using corresponding local features from their neighborhood. The

methodology enhances prior approaches in point pattern based registration, like RANdom SAmple Consensus (RANSAC) and Thin Plate Spline-Robust Point Matching (TPS-RPM), to enable registration of cell phone and camera captured documents under non-rigid transformations. Three novel aspects are embedded into the methodology: (i) histogram based uniformly transformed correspondence estimation, (ii) clustering of points located near the ROI to select only close by regions for matching, and (iii) validation of the registration in RANSAC and TPS-RPM algorithms. Experimental results on a dataset of 480 images captured using iPhone 3GS and Logitech webcam Pro 9000 have shown an average registration accuracy of 92.75% using Scale Invariant Feature Transform (SIFT).

Robust local features for logo identification are determined empirically by comparisons among SIFT, Speeded-Up Robust Features (SURF), Hessian-Affine, Harris-Affine, and Maximally Stable Extremal Regions (MSER). Two different matching methods are presented for categorization: matching all features extracted from the query document as a single set and a segment-wise matching of query document features using segmentation achieved by grouping area under intersecting dense local affine covariant regions. The later approach not only gives an approximate location of predicted logo classes in the query document but also helps to increase the prediction accuracies. In order to facilitate scalability to large data sets, inverted indexing of logo class features has been incorporated in both approaches. Experimental results on a dataset of real camera captured documents have shown a peak 13.25% increase in the F–measure accuracy using the later approach as compared to the former.

REGISTRATION AND CATEGORIZATION OF CAMERA CAPTURED DOCUMENTS

by
Venkata Gopal Edupuganti

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computer Science

May 2012

APPROVAL PAGE

REGISTRATION AND CATEGORIZATION OF CAMERA CAPTURED DOCUMENTS

Venkata Gopal Edupuganti

Dr. Frank Y. Shih, Dissertation Advisor — Date
Professor of Computer Science, NJIT

Dr. James McHugh, Committee Member — Date
Professor of Computer Science, NJIT

Dr. Dimitri Theodoratoes, Committee Member — Date
Associate Professor of Computer Science, NJIT

Dr. Vincent Oria, Committee Member — Date
Associate Professor of Computer Science, NJIT

Dr. Suryaprakash Kompalli, Committee Member — Date
Research Scientist, Hewlett-Packard Labs, Bangalore

BIOGRAPHICAL SKETCH

Author: Venkata Gopal Edupuganti

Degree: Doctor of Philosophy

Date: May 2012

Undergraduate and Graduate Education:

- PhD in Computer Science,
  New Jersey Institute of Technology, Newark, NJ, 2012
- Masters in Information Technology,
  University of Hyderabad, Hyderabad, India, 2007
- Bachelors in Information Technology,
  JNTU, Hyderabad, India, 2004

Major: Computer Science

Presentations and Publications:

V. G. Edupuganti, F. Y. Shih, and S. Kompalli. Logo detection driven categorization of camera captured documents using SIFT. *Pattern Recognition*, 2012. (under submission)

V. G. Edupuganti, F. Y. Shih, and S. Kompalli. Categorization of camera captured documents based on logo identification. In *Proceedings of the Fourteenth International Conference on Computer Analysis of Images and Patterns - Volume Part II*, pages 130–137, Seville, Spain, August 2011.

V. G. Edupuganti, V. A. Agarwal, and S. Kompalli. Registration of camera captured documents under non-rigid deformation. In *Proceedings of the Twenty Fourth IEEE International Conference on Computer Vision and Pattern Recognition*, pages 385–392, Colorado Springs, CO, June 2011.

V. G. Edupuganti, F. Y. Shih, and I. Chang. An efficient block-based fragile watermarking system for tamper localization and recovery. *Intelligent Automation and Soft Computing*, 17(2):257–267, 2011.

V. G. Edupuganti and F. Y. Shih. Authentication of JPEG images based on genetic algorithms. *The Open Artificial Intelligence Journal*, 4:30–36, 2010.

V. G. Edupuganti, V. N. K. P. Munnangi, and R. Vadlamani, Fast and accurate watermark retrieval using evolutionary algorithms. *International Journal of Computer Information Systems and Industrial Management Applications*, 2:121–136, 2010.

F. Y. Shih and V. G. Edupuganti. A differential evolution based algorithm for breaking the visual steganalytic system. *Soft Computing*, 13(4):345–353, 2009.

V. G. Edupuganti, V. N. K. P. Munnangi, and R. Vadlamani, Evolutionary algorithms for fast and accurate watermark retrieval. In *Proceedings of the IEEE World Congress on Nature & Biologically Inspired Computing*, pages 991–1004, Coimbatore, India, December 2009.

V. G. Edupuganti, R. Vadlamani, and V. N. K. P. Munnangi, Efficient watermark retrieval through hopfield neural network. *Soft Computing: New Research*, Nova Publishers, New York, 2008.

To My Late Mother

ACKNOWLEDGMENT

I owe my deep gratitude to all those who have made everything possible to complete this dissertation. First and foremost, my sincere thanks goes to my dissertation advisor, Dr. Frank Y. Shih for his endless support, encouragement, and precious advises during my graduate studies. I am grateful to Dr. James McHugh, Dr. Vincent Oria, Dr. Dimitri Theodoratoes, and Dr. Suryaprakash Kompalli for serving in my dissertation committee as well as their encouragement.

I take this opportunity to thank all my research internship mentors, Dr. Shin'ichi Satoh, National Institute of Informatics, Tokyo; Dr. Suryaprakash Kompalli, Research Scientist, Hewlett-Packard Labs, Bangalore, India; and Dr. Vivek Kwatra, Research Scientist, Google, Mountain View, CA for introducing me to new areas of research and involving me in cutting edge product development. My special thanks goes to Dr. Suryaprakash Kompalli for his help and motivation in shaping up this dissertation.

I would also like to thank my friends at NJIT, Shuo Chen, Chandralekha De, Zhimming Liu, and Jichao Sun for their generous help in making my doctoral studies memorable. Additionally, I thank my friends, Gowtham Atluri, PhD Candidate, University of Minnesota and Saigopal Thota, PhD Candidate, University of California, Berkeley, for their valuable discussions.

Finally, I would like to thank Dr. David Nassimi and Dr. George Olsen for being flexible in awarding my teaching assistantship.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter Page

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES
(Continued)

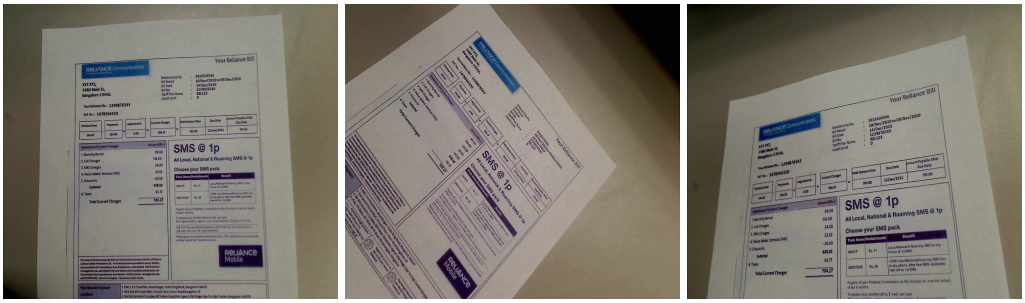
LIST OF FIGURES
(Continued)

Figure **Page**

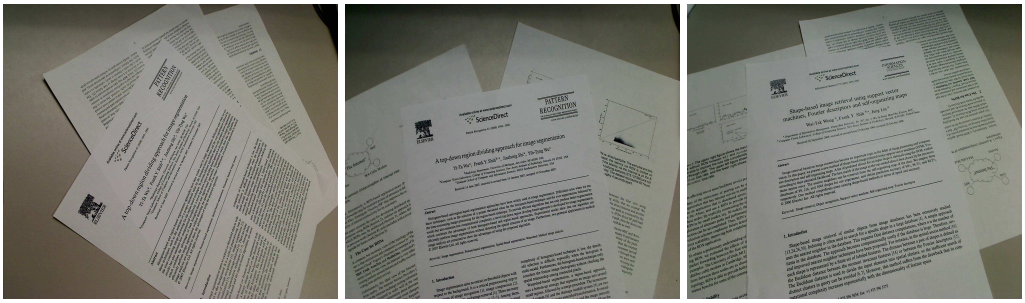
CHAPTER 1

INTRODUCTION

Camera captured document image analysis [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15] concerns with processing of documents captured with hand-held sensors, mobile phones incorporated with cameras, or other capturing devices using advanced image processing, computer vision, pattern recognition, and machine learning techniques. As there is no constrained capturing in the real world, the captured documents suffer from illumination variation, viewpoint variation, highly variable scale/resolution, background clutter, occlusion, and complex non-rigid deformations i.e., folds and crumples being seen in paper images. Figure 1.1 shows a glimpse of camera captured documents with respect to highly varying challenging conditions. In recent years, few techniques have been developed in flattening of curved documents [16, 5, 8], document mosaicing [4, 7, 9, 12, 15], curled text-line segmentation [1], robust text extraction from images [2, 14] and document image retrieval [17, 18, 10, 11, 13]. Document registration is a problem where the image of a template document whose layout is known is registered with a test document image. Literature in document mosaicing addressed registration of captured documents with the assumption of considerable amount of single chunk overlapping content. These methods can not be directly applied to registration of forms, bills, and other commercial documents where the fixed content is distributed into tiny portions across the document. Literature in document image retrieval addressed categorization of documents based on text, figures, etc. However, the scalability of existing document categorization methodologies based on logo identification is very limited [19, 20, 21, 22, 23]. This dissertation focuses on two problems (i) registration of captured documents where the overlapping content is distributed into tiny portions across the documents and (ii) categorization of captured documents into predefined logo classes that scale to large datasets using local invariant features.



(a) Illumination and view-point variation



(b) Background clutter

(c) Occlusion

(d) Folds and crumples

Figure 1.1 A glimpse of camera captured document images.

1.1 Local Features

This section introduces a set of local invariant features [24, 25, 26, 27, 28, 29, 30] used in this dissertation. Due to robustness to local changes e.g., illumination and view-point in images, local features have been drawing more attention of researchers compared to global features e.g., color histograms and texture features [31] in a wide variety of tasks including image classification [32, 33], image search [34, 35], video copy detection [36, 37], robust text detection in document images [2], and so on. Features considered are Scale Invariant Feature Transform (SIFT) [27, 28], Speeded-Up Robust Features (SURF) [24, 25], Harris-Affine regions [26, 30], Hessian-Affine regions [30], and Maximally Stable Extremal Regions (MSER) [38, 30]. Figure 1.2 shows different local features extracted from an example image. The following subsections briefly describe each feature.

1.1.1 Scale Invariant Feature Transform (SIFT)

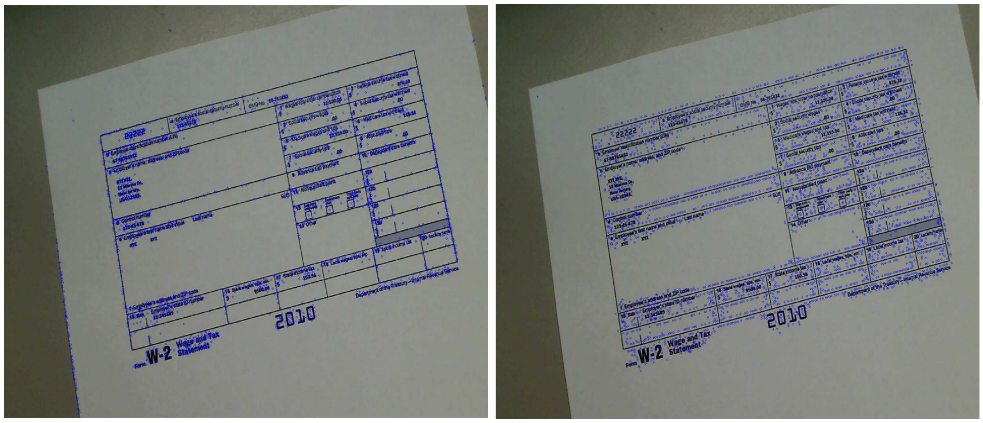
SIFT [39, 40, 41, 27, 28] features are shown invariant to image scale and rotation, and are robust to substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination. Feature extraction involves a four stage cascade filtering approach, in which most expensive operations are done at locations that pass initial test. The following steps comprise SIFT feature extraction:

i Scale-Space Extrema Detection: Construct scale space [42, 27, 28] $L(x, y, \sigma)$ of an image from the convolution of a variable-scale Gaussian, $G(x, y, \sigma)$ with an input image $I(x, y)$.

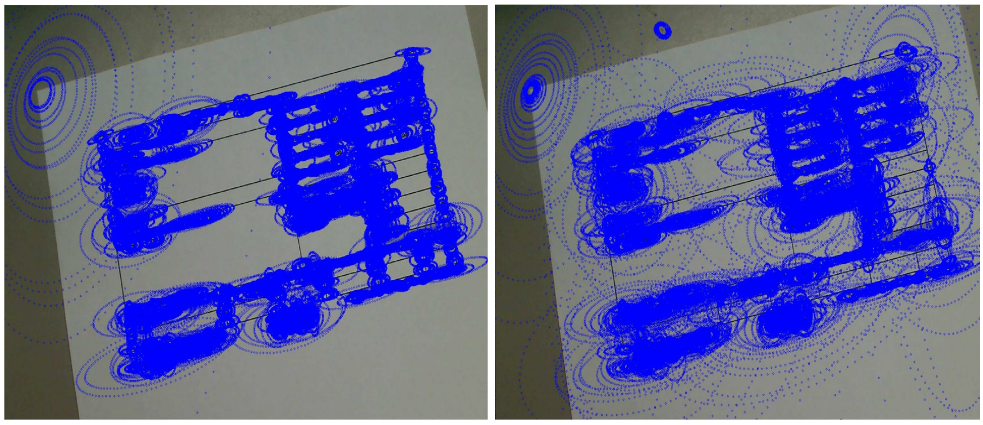
$$L(x, y, \sigma) \quad = \quad G(x, y, \sigma) * I(x, y) \tag{1.1}$$

where * is the convolution operation in x and y, and

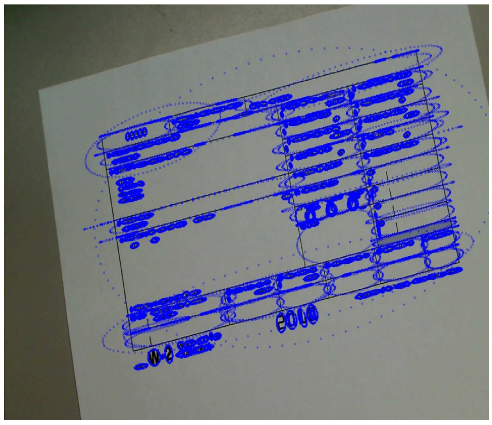
$$G(x, y, \sigma) \quad = \quad \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{1.2}$$



(a) SIFT (b) SURF



(c) Hessian-Affine (d) Harris-Affine



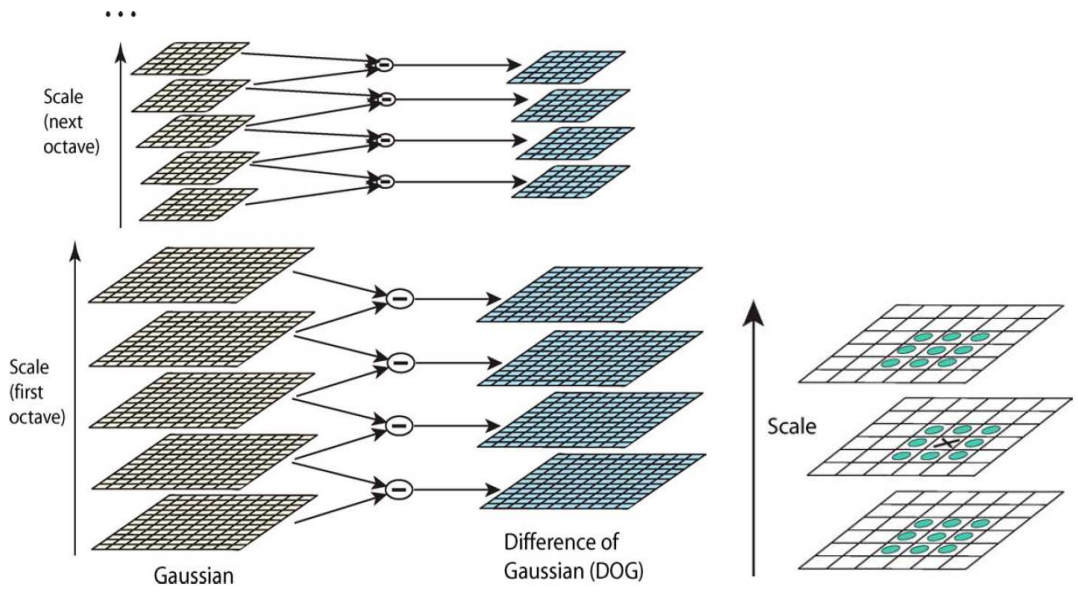
(e) MSER

Figure 1.2 Local features example.

Compute difference of Gaussian function $D(x, y, \sigma)$ from the difference of two near by scales separated by a constant multiplicative factor k:

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned} \tag{1.3}$$

The difference of Gaussian function is a close approximation of scale-normalized Laplacian of Gaussian [42, 27, 28]. Divide each octave i.e., doubling of σ of scale space into an integer number s of intervals, $k = 2^{1/s}$. Compare each sample point in $D(x, y, \sigma)$ to its eight neighbors in the current image and nine neighbors in the scale above and below, consider it as a candidate keypoint for further investigation in later stages only if it is either maximum or minimum of all its neighbors as shown in Figure 1.3.



(a) Scale space construction (b) keypoint identification

Figure 1.3 Scale space extrema computation [28].

ii Keypoint Localization: Fit candidate keypoints to nearby data for location, scale, and ratio of principle curvatures. This is achieved by Taylor expansion [27, 28] (up

to quadratic terms) of the scale-space function, $D(x, y, \sigma)$, shifted so that the origin is at the keypoint:

$$D(x) = D + \frac{\partial D^T}{\partial x}x + \frac{1}{2}x^T\frac{\partial^2 D}{\partial x^2}x \tag{1.4}$$

where D and its derivatives are evaluated at candidate keypoint and $x = (x, y, \sigma)^T$ is the offset from this point. The location of the extremum is determined by taking the derivative of this function with respect to x and setting it to zero, giving

$$\hat{x} = -\frac{\partial^2 D^{-1}}{\partial x^2}\frac{\partial D}{\partial x} \tag{1.5}$$

Reject keypoints with low contrast i.e., function value at the extremum $D(\hat{x})$ less than 0.03 assuming image pixel values in the range [0,1] and ratio between principle curvatures greater than ten.

$$D(\hat{x}) = D + \frac{1}{2}\frac{\partial D^T}{\partial x}\hat{x} \tag{1.6}$$

iii Orientation Assignment: Select the Gaussian smoothed image L closest to scale of the keypoint. Compute an orientation histogram of 36 bins covering 360 degree range of orientations from the gradient orientations $\theta(x, y)$ of sample points within a region around the keypoint. Each sample added to the histogram is weighted by its gradient magnitude $m(x, y)$ and by a Gaussian-weighted circular window with a σ that is 1.5 times that of the scale of the keypoint. Peaks in the orientation histogram correspond to dominant orientations of local gradients. The highest peak and the peaks within 80% of the highest peak create keypoints with corresponding orientation. Fit a parabola to three histogram values closest to each peak to interpolate the peak position.

$$\begin{aligned} m(x, y) &= \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \\ \theta(x, y) &= tan^{-1}((L(x, y+1) - L(x, y-1))/L(x+1, y) - L(x-1, y))) \end{aligned} \tag{1.7}$$

iv Keypoint Descriptor: Sample gradient magnitudes and orientations around keypoint in Gaussian blur image closest to the scale of the keypoint. Rotate coordinates of descriptor and gradient orientations relative to the keypoint orientation. Assign weight to gradient magnitudes using a Gaussian function with σ equal to 1.5 times width of the descriptor window. Assign gradient magnitudes and orientations to a 4×4 array of eight bin orientation histograms i.e., divide sampling region around the keypoint into 4×4 subgrids using trilinear interpolation, which leads to a $4\times4\times8$ = 128 dimensional descriptor (Figure 1.4 shows computation of descriptor in 2×2 subgrids). Normalize the descriptor to unit length in order to be invariant to illumination changes and threshold bin values greater than 0.2.

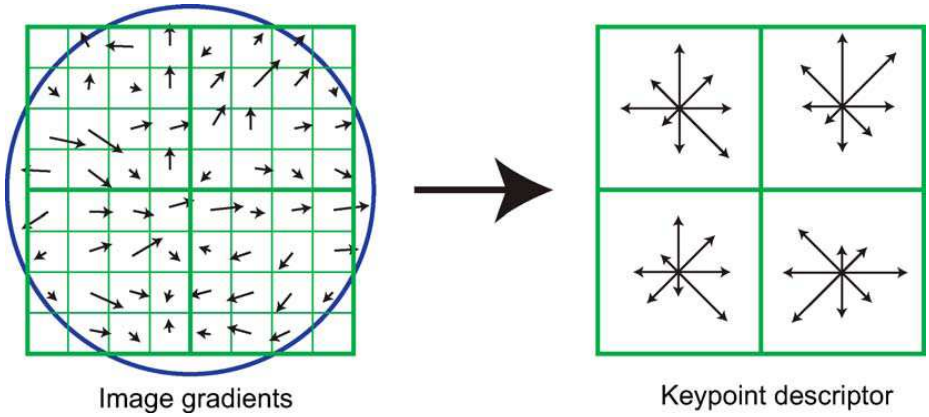


Figure 1.4 A 2×2 array of SIFT description (right) from an 8×8 array of samples (left) [28].

1.1.2 Speeded-Up Robust Features (SURF)

SURF [24, 25, 43] is invariant to scale and in-plane rotations, and provide some degree of robustness to skew, anisotropic scaling, and perspective effects. The primary focus of SURF is on fast detection of interest points in scale space. This is achieved by using integral images for fast computation of box type convolution filters. Unlike SIFT, interest points

are detected at the extrema of determinant of Hessian matrix $H(x, y, \sigma)$.

$$H(x, y, \sigma) = \begin{bmatrix} L_{xx}(x, y, \sigma) & L_{xy}(x, y, \sigma) \\ L_{xy}(x, y, \sigma) & L_{yy}(x, y, \sigma) \end{bmatrix} \tag{1.8}$$

where $L_{xx}(x, y, \sigma)$ is the convolution of the Gaussian second order derivative $\frac{\partial^2}{\partial x^2} G(x, y, \sigma)$ with the image I in point (x, y) at scale σ, and similarly for $L_{xy}(x, y, \sigma)$ and $L_{yy}(x, y, \sigma)$. This kind of interest point detection favors blob-like structures. In order to assign orientation to detected interest points, Haar wavelet responses are computed in x and y direction within a circular neighborhood of radius $6s$ around the interest point, where s denote the scale at which the interest point was detected. The wavelet responses are further weighted with a Gaussian $\sigma = 2s$ centered at the interest point. A sliding window approach is used to determine the dominant orientation by distributing horizontal responses along abscissa and vertical responses along ordinate. For the extraction of descriptor, a square region with size $20s$ centered around the interest point is selected and oriented along the dominant orientation. The region is split up regularly into 4×4 square subregions and each subregion is described by $(\sum d_x, \sum |d_x|, \sum d_y, \sum |d_y|)$, where d_x and d_y denote Haar wavelet response in horizontal and vertical direction, respectively (Figure 1.5 shows computation of descriptor in 2×2 subgrids). This kind of description leads to a 4×4×4 = 64 dimensional descriptor. A 128 dimensional description is achieved by following extended description [24].

1.1.3 Harris-Affine Regions

Harris-Affine regions [26, 44, 30] are based on affine normalization around Harris points. Interest points are selected in scale space using second moment matrix of intensity gradient i.e., autocorrelation matrix.

$$M = \sigma_D^2 G(x, y, \sigma_I) * \begin{bmatrix} I_x^2(x, y, \sigma_D) & I_x I_y(x, y, \sigma_D) \\ I_x I_y(x, y, \sigma_D) & I_y^2(x, y, \sigma_D) \end{bmatrix} \tag{1.9}$$

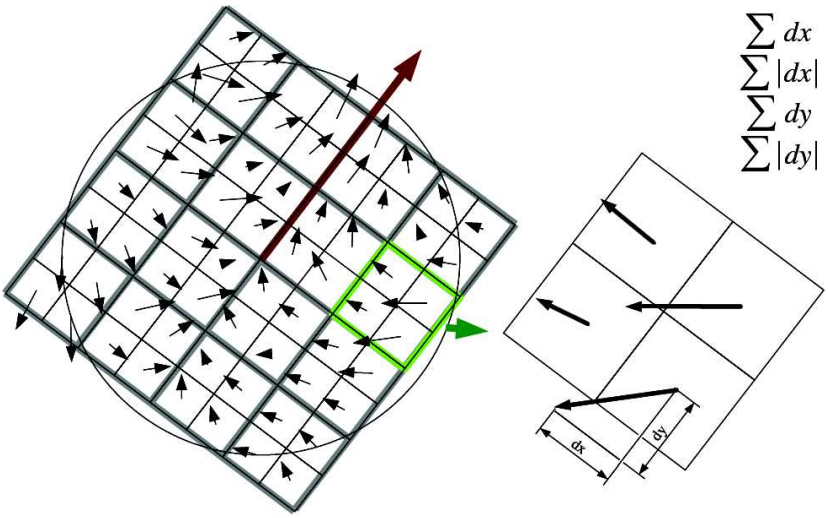


Figure 1.5 A 2×2 array of SURF description (right) from an 8×8 array of samples (left) [25].

The local image derivatives are computed with Gaussian kernels of scale σ_D (differentiation scale) and then averaged in the neighborhood of the point by smoothing with a Gaussian window of scale σ_I (integration scale). The eigen values of the matrix represent two principal changes in a neighborhood of the point, and interest points are selected at the points in which the signal change is significant in orthogonal directions. An iterative estimation of elliptical affine region [45, 46] around interest points is performed using autocorrelation matrix. Each elliptical region is described by using SIFT description.

1.1.4 Hessian-Affine Regions

Hessian-affine regions [30] are defined by affine normalization around Hessian points. Interest points are selected in scale space using Hessian matrix.

$$H = \begin{bmatrix} I_{xx}(x, y, \sigma_D) & I_{xy}(x, y, \sigma_D) \\ I_{xy}(x, y, \sigma_D) & I_{yy}(x, y, \sigma_D) \end{bmatrix} \tag{1.10}$$

The local maximum of the determinant of the matrix defined by second derivatives is used to select the interest points. Similar to Harris-Affine regions, an iterative estimation of

elliptical affine region [45, 46] around interest points is performed using autocorrelation matrix and the corresponding region is described by following SIFT description.

1.1.5 Maximally Stable Extremal Regions (MSER)

A Maximally Stable Extremal Region (MSER) [47, 38, 30] is a connected component of an appropriately thresholded image. All the pixels inside MSER have either higher or lower intensity than all the pixels on its outer boundary. To extract MSER from an image, first, pixels are sorted by intensity. After sorting, pixels are marked in the image (either in decreasing or increasing order) and the list of growing and merging connected components and their areas is maintained using the union-find algorithm. During this process, the area of each connected component as a function of intensity is stored. The maximally stable ones are those corresponding to thresholds where the relative area change as a function of relative change of threshold is at a local minimum i.e., MSER are the parts of the image where local binarization is stable over a large range of thresholds. Finally, each MSER is approximated by an affine invariant ellipse and described using SIFT.

1.2 Feature Matching

Local features discussed in the previous section help to represent an image invariant to illumination, scale, and view-point. Feature matching plays a vital role to effectively use extracted features to perform image registration [48, 49], search [18], retrieval [34, 35], and so on. The first step in matching is to establish point correspondences between two images. A correspondence is established between similar features in two different images based on a distance between feature vectors e.g., Euclidean distance or Mahalanobis distance [50]. Lowe [28] proposed an efficient way of establishing correspondences by exploring nearest neighbor distances in descriptor space. Due to noise in feature extraction and description, the established point correspondences contain outliers. In order to perform high-level image processing tasks such as registration and retrieval,

robust outlier elimination methodologies should be adapted. Several outlier elimination techniques have been proposed in literature, such as least squares minimization [51], RANdom SAmple Consensus (RANSAC) [52, 51, 53, 54], Robust Point Matching [55, 56], Hough clustering [57, 58, 51, 28], and so on. On the other hand, inverted indexing [34] is a very popular technique to conduct matching in large scale. The following subsections briefly describe popular outlier eliminations mechanisms and inverted index:

1.2.1 Least Squares Minimization

Let x_i and $y_i, 1 \leq i \leq K$ be the two corresponding point sets obtained by establishing correspondences using feature similarity measure e.g., Euclidean distance. It computes optimal transformation A by minimizing the following transformation error [51]:

$$A^* = arg\min_{A} \sum_{i=1}^{K} ||y_i - Ax_i||^2 \tag{1.11}$$

The size of transformation matrix A is 2×3 for affine transformation (6 degrees of freedom) and 3×3 for perspective transformation (8 degrees of freedom).

1.2.2 Hough Transform Clustering

Hough transform [51] method follows the principle of maximum likelihood estimation. It maps the data into quantized parameter space and seeks for the most likely parameter values to interpret the data through clustering. The number of parameters to estimate are 6 and 8 for affine and perspective transformations respectively. As each correspondence (x_i, y_i) can be represented by one or more transformations, it vote for all the relevant underlying transformation parameters in the quantized space. The optimal transformation parameters correspond to the bin that accumulates most number of votes.

1.2.3 RANdom SAmple Consensus (RANSAC)

RANSAC [52, 51] iteratively estimates parameters of an underlying transformation model from a set of observed data i.e., correspondences which contains outliers (correspondences that do not fit to the model). It is a non deterministic algorithm that produces probabilistic accuracy, with the accuracy increasing as more iterations are conducted. The key assumption is that the data consists of inliers i.e., correspondences that fit to the transformation model. Several variants of RANSAC [52] have been proposed so far with different objectives such as accuracy, speed, and robustness. The input to the RANSAC algorithm is point correspondences, underlying transformation model e.g., affine and some confidence parameters e.g., accuracy. RANSAC iterations are divided into two stages:

i Hypothesis Generation: Randomly select a subset of point correspondences and estimate assumed transformation model with the chosen subset.

ii Hypothesis Evaluation: Test entire point correspondences against the estimated model. Divide the point correspondences into hypothetical inliers (that agree with the estimated model) and hypothetical outliers (that do not agree with the estimated model). Update the best estimated model so far with the current model if the current model has more number of hypothetical inliers.

The algorithm iterates until a fixed number of iterations, or predefined accuracy, or a combination of both. Reestimate the retrieved model using only inliers. The performance of RANSAC degrades with increase in the number of outliers in the point correspondences.

1.2.4 Thin Plate Spline-Robust Point Matching (TPS-RPM)

Thin Plate Spline-Robust Point Matching (TPS-RPM) [55, 56] algorithm is designed to derive underlying non-rigid transformation function from point correspondences containing outliers. The reason behind choosing thin plate spline [59] in robust point matching framework is that it is the only spline that can be easily decomposed into

affine and non-affine subspaces while minimizing a bending energy based on the second derivative of the spatial mapping. Robust point matching is similar to Expectation Maximization (EM) algorithm, which iteratively minimizes the following least squares energy function using deterministic annealing and soft-assignment:

$$E_{TPS}(f) = \sum_{a=1}^{K} ||y_a - f(v_a)||^2 + \lambda \int\int [(\frac{\partial^2 f}{\partial x^2})^2 + 2(\frac{\partial^2 f}{\partial x \partial y})^2 + (\frac{\partial^2 f}{\partial y^2})^2] dxdy \quad (1.12)$$

where f is a thin plate spline mapping function between corresponding point sets y_a and v_a and λ is a regularization parameter.

$$f(v_a, d, w) = v_a.d + \phi(v_a).w \quad (1.13)$$

where d is a $(D+1) \times (D+1)$ matrix representing affine transformation (D is the dimension of points), w is a $K \times (D+1)$ warping coefficient matrix representing non-affine deformation (K is the first point set cardinality), and $\phi(v_a) = ||v_b - v_a||^2 log||v_b - v_a||$ is a $1 \times K$ TPS kernel.

1.2.5 Inverted Indexing

The methodologies introduced in previous subsections are very expensive to perform matching in large scale such as similar image search and retrieval, where a set of features extracted from a query image i.e., image under observation are matched against a set of features from all the images in a dataset. Inverted indexing [60, 61, 62] is a widely used technique in text retrieval and has been drawing more attention of researchers to conduct large scale visual search and retrieval [34, 63, 35]. The first step involved in inverted indexing is the vector quantization of feature descriptors into visual words. Unsupervised clustering methods such as K-means [64] and hierarchical K-means [64] are generally used to compute clusters of local feature descriptor vectors, and each cluster centroid is termed as a visual word. The set of all visual words comprises to visual word vocabulary. An inverted file is structured like an ideal book index. It has an entry for each visual word in

the vocabulary followed by a list of all the images (possibly position in the image) in which the visual word occurs. A query image is represented as a vector of visual word frequencies. The precomputed inverted file is parsed with the query image visual words and images that have sufficient number of visual words in intersection with the query image visual words will be retrieved. The significance of this approach comes from the fact that only those images that have visual words in common with query image are retrieved. There exists a number of efficient ways in using inverted files such as tf-idf weighting [63], Hamming distance [34], bundling features [35], and so on.

1.3 Topics Overview

Image registration [65, 66, 67, 48, 68, 69, 70, 71, 49, 72, 73, 74, 75, 76, 77] is a very well known problem in image processing and computer vision, which derives a geometric transformation between an arbitrarily deformed image and a known image. Few techniques have been developed for camera captured document image registration [4, 7, 9, 12]. These methods are limited to affine and skew transformations and assume that the content is fixed between template and test images. Chapter 2 presents a novel methodology to register Regions Of Interest (ROI) under complex non-rigid deformations. Why only registration of ROI? Applying traditional Optical Character Recognition (OCR) [78, 79, 80] methods on entire document gives a lot of noise due to camera capturing deformations. However, applying OCR on a region that closely contains ROI extracts the content in ROI more accurately. Figure 1.6 shows the result of applying OCR on various region sizes containing social security number (SSN) in the W2 form. Figures 1.6(a) and 1.6(c) show bigger regions containing SSN, and the corresponding OCR results are shown in Figures 1.6(b) and 1.6(d), respectively. Figures 1.6(e) and 1.6(e) show that the application of OCR on a region that closely contains ROI yields more accurate results. The key point here is deriving a non-rigid transformation function that maps a user defined ROI in template image (i.e., reference image) to the captured image. This is achieved by providing enhancements to the

previous literature in RANSAC [52, 51] and TPS-RPM [55, 56] algorithms. Experimental results on a dataset of 480 images captured using iPhone 3GS and Logitech webcam Pro 9000 have shown an average registration accuracy of 92.75% using SIFT.

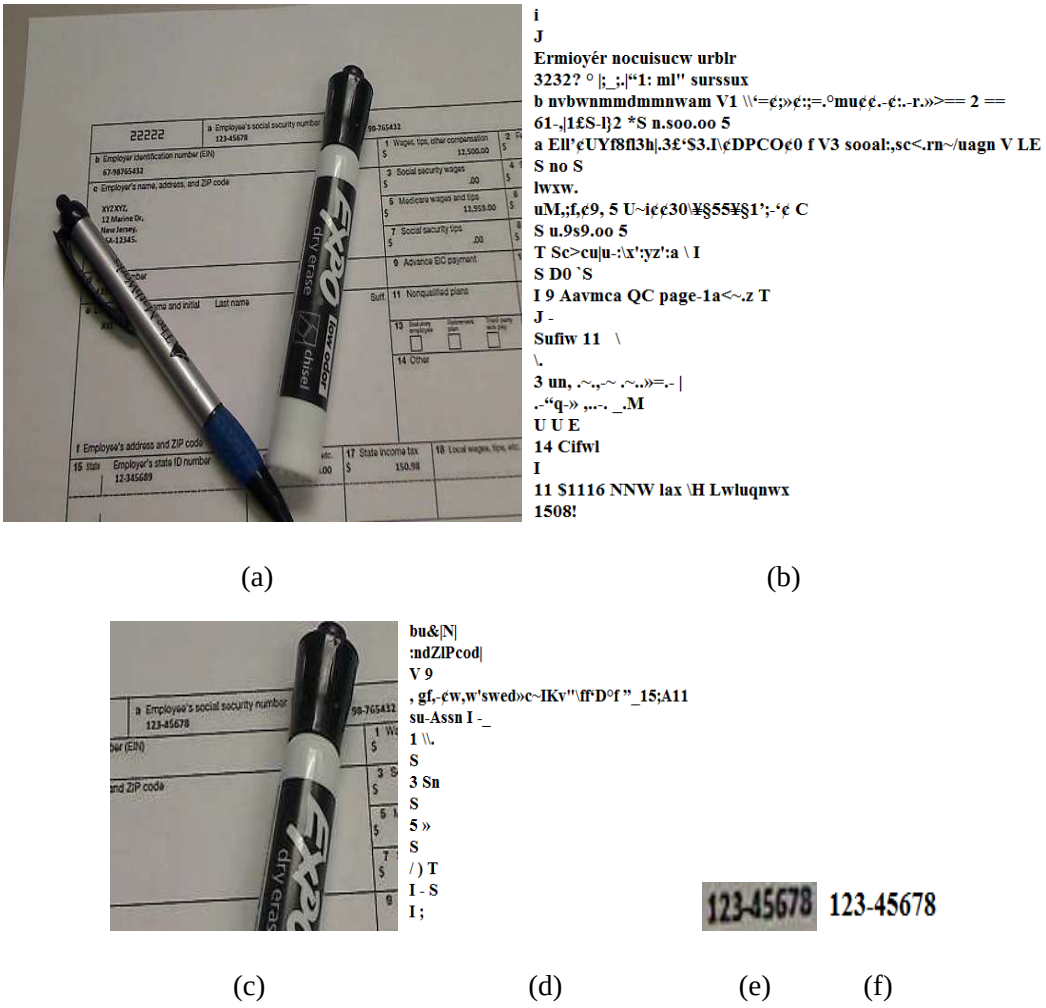


Figure 1.6 Application of OCR on different region sizes of camera captured document.

Chapter 3 presents a methodology to categorize camera captured documents into predefined logo classes. Existing approaches in logo detection [19, 20, 21, 22, 23] are limited to scanned documents and few approaches have addressed logo detection in natural scenes [81, 82]. Literature in document retrieval [17, 83, 18, 10, 11, 12] requires either full or partial archived document as a query rather than just logo. A signature detection and matching methodology is presented in [84]. Due to large diversity in logos i.e., text,

graphics, and a mixture of both and the noise introduced by camera capturing i.e., scale, illumination, and viewpoint variations, detection of logos is very challenging. Robust local features are derived by comparisons among various local affine invariant features under different criteria such as feature repeatability, distinctiveness, etc. A two step matching methodology is presented to efficiently search and retrieve the underlying logo classes. Besides, Hamming Embedding [85, 34] is applied to reduce the noise in descriptor quantization and inverted indexing of logo classes to conduct real time category prediction. Experimental results on a data set of real camera captured documents have shown the behavior of different feature representations in category prediction.

Chapter 4 presents a segment-wise matching approach to perform document categorization by detecting logos. Literature in block segmentation of documents addressed the segmentation of printed [86] and scanned documents [87, 88, 89, 90]. The key step involved in such approaches is the binarizaion, which introduces a lot of noise in camera captured documents. In order to overcome camera capturing artifacts, an approach to segment query document image is presented by grouping area under intersecting dense local affine covariant regions [30]. The presented methodology not only improves the prediction accuracies but also gives an approximate position of the predicted logo classes in the query document. Experimental results on a data set of real camera captured documents have shown a peak 13.25% increase in the F-measure [64] accuracy as compared to the methodology presented in Chapter 3.

The conclusions and future work are presented in Chapter 5, where the major contribution of this dissertation is summarized and future research directions are discussed.

CHAPTER 2

REGISTRATION OF REGIONS OF INTEREST

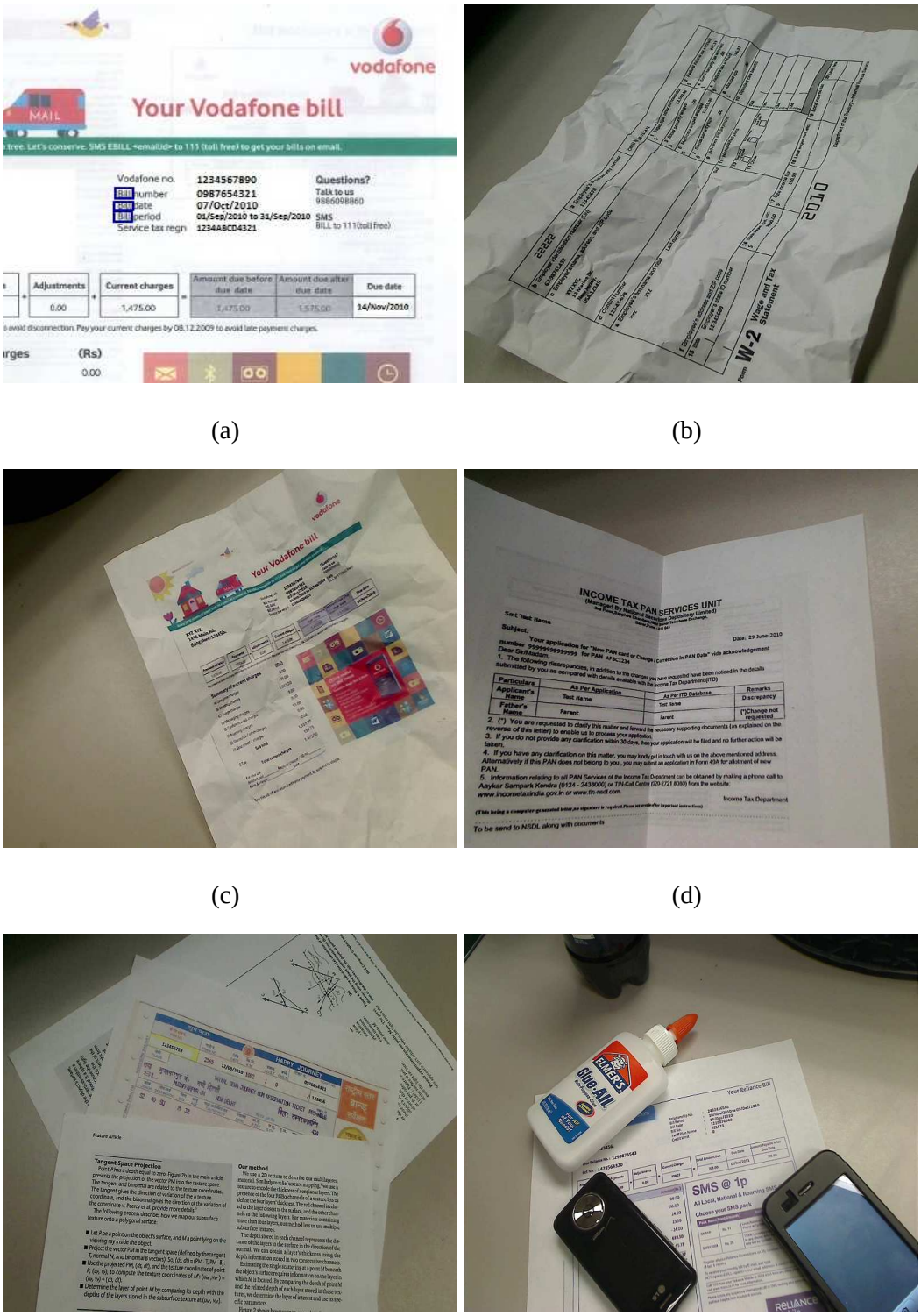
Document registration [4, 7, 9, 12] is a problem where the image of a template document whose layout is known is registered with a test document image. Given the registration parameters, layout of the template image is superimposed on the test document. Registration algorithms have been popular in applications, such as forms processing where the superimposed layout is used to extract relevant fields. The proliferation of camera captured images makes it necessary to address camera noise such as non-uniform lighting, clutter and highly variable scale/resolution along with complex non-rigid deformations such as folds and crumples. This chapter presents a novel registration methodology for user defined Regions of Interest (ROI) under complex deformations using enhancements to prior approaches in point pattern based registration, like RANdom SAmple Consensus (RANSAC) and Thin Plate Spline-Robust Point Matching (TPS-RPM). Three significant aspects that comprise the framework are (i) histogram based uniformly transformed correspondence estimation, (ii) clustering of points located near ROI to select only close by regions for matching, and (iii) validation of the registration in RANSAC and TPS-RPM algorithms. Experimental results section discusses behavior of registration accuracies using SIFT and SURF features.

2.1 Related Work

Image registration [55, 56, 91, 92, 93, 94] by establishing correspondences across interest points in image pairs has been well studied in image processing and computer vision. Registration becomes more challenging when outliers exist in the correspondence set. These outliers could arise from noise in image acquisition, feature extraction, and/or matching. Several local invariant (e.g., scale, affine, and intensity) detectors and

descriptors [25, 95, 28, 29, 30, 96] have been proposed to overcome the natural variations in image acquisition. Feature similarity measures, such as L2 norm, cosine distance, etc, together with outlier elimination techniques, such as RANdom Sample Consensus (RANSAC) [52], Hough Transform [28], and TPS-RPM [55, 56], have been applied to establish true correspondences. The goal of most techniques is to estimate underlying transformation function across natural images for purposes such as image stitching [97, 98], image augmentation [99], or camera geometry estimation [54].

Camera captured documents differ from natural images in a few critical areas: (i) Non-linear deformations, such as folds, are common in documents, (ii) In most forms, the filled values are large in number and the amount of similar content between the test and template is a percentage of the document content, and (iii) Content such as logos, and even text is repeated at multiple locations within the document. Figure 2.1 shows a few forms that are quite sparse in content, where the same text occurs at multiple locations in a document. This results in a new class of outliers that are similar in the domain of local features but correspond to a different region that is not aligned with the global image layout. The existence of correspondences from one region to multiple regions increases the number of outliers and has an adverse effect on traditional iterative methods such as RANSAC. These challenges are in addition to known problems in camera capture, such as lighting variations, clutter, camera equipment differences, and scale. Document image processing has earlier used registration techniques for forms processing. The motivation has frequently been that information from a small part of the document is critical for most user applications. For example, the amount and date on a receipt is all that is needed as an input to a tax software. Limiting downstream processing to relevant regions is known to be useful both from the view of accuracy and speed. To extract only relevant regions, a test image is registered with a template image that has known layout. Here, selected regions of text are extracted from a filled in form (test image) using information about the form layout (template image). Registration parameters are used to overlay the layout of the template



(a) (b)

(c) (d)

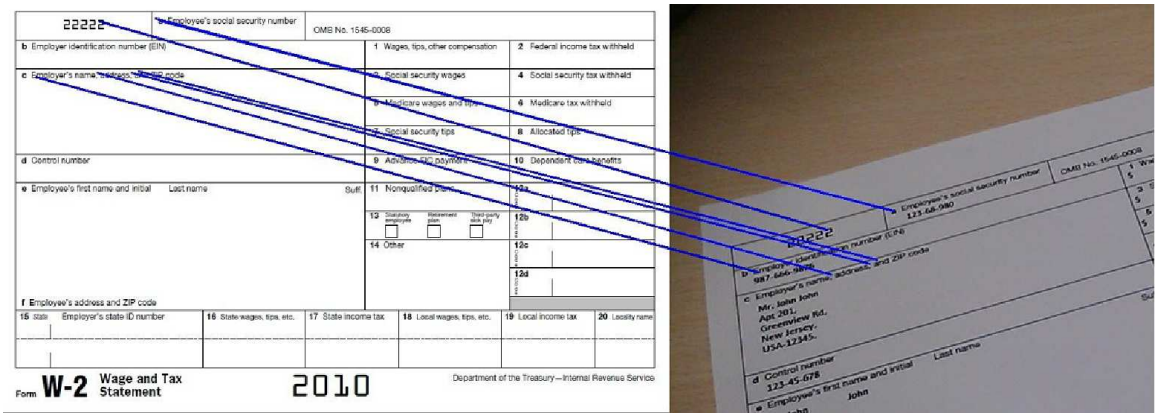
(e) (f)

Figure 2.1 (a) Regions in blue rectangle are similar i.e., ”Bill”, (b)-(f) captured document images with non-planar deformations and occlusion.

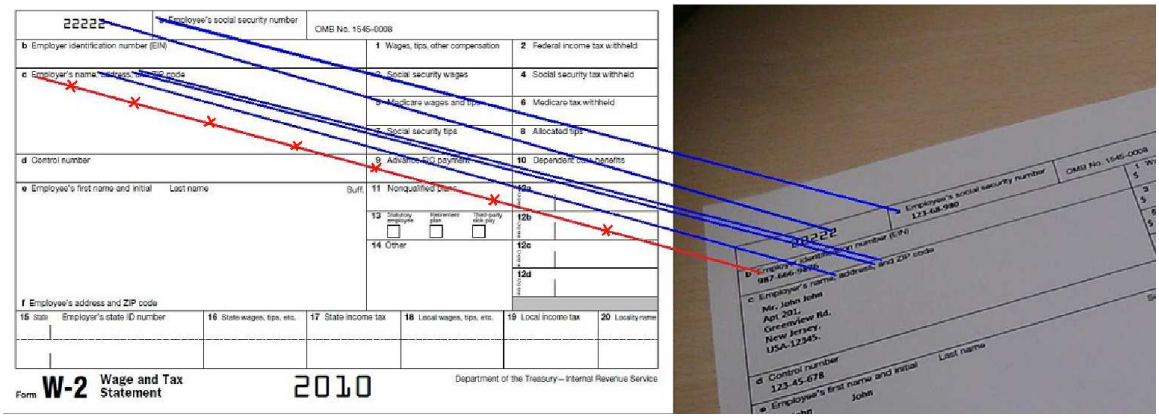
image onto the test image. The layout specifies geometric positions of the relevant fields, which are then extracted from the test image. While several prior techniques [100, 101, 102, 103] fall in this area, these methods address only affine transformations and assume a high quality image.

Approaches such as RANSAC [52] and Hough clustering [28] estimate true correspondence by fitting a transformation function to existing correspondences. They are designed to eliminate correspondences between points that have high feature similarity but do not agree with the global image geometry. By design, outliers would also influence the transformation function, and would be considered as inliers if they conform to the underlying transformation. For example, Figure 2.2 shows an outlier that conforms to global geometry has been considered as an inlier. This, and similar outliers deviate the region of interest from the desired location. Applying these methods directly for non-rigid registration is not acceptable as the underlying transformation function varies at different parts of the image. Several non-rigid registration frameworks [55, 56, 92, 93, 94] have been developed for the non-rigid registration of medical images. One of them is the Robust Point Matching (RPM) [56, 94] algorithm, which formulates the registration problem as a maximum likelihood estimation problem using mixture models. Chui and Rangarajan [56] embedded the Expectation Maximization (EM) frame work in a deterministic annealing scheme by considering the soft-assignment of point sets to allow partial matches. Thin Plate Splines (TPS) [59] are used for the estimation of underlying transformation function, as TPS can be decomposed into affine and non-affine sub spaces. The Robust Hybrid Deformable Matching (RHDM) framework [94] incorporates feature dissimilarity measure into the TPS-RPM framework. Sofka et al. [93] pointed out that the TPS-RPM algorithm would fail on extraneous structures, such as H-shape point sets, as it tries to align the center of mass of the point sets in the early iterations of the algorithm, leading to a bias in the estimate, which it can not overcome in the later stages of the algorithm. Recently, Myronenko et. al. [92] incorporated motion coherence theory into the

framework in the place of TPS. All these methods fail to consider a few factors: (i) Initial correspondences are not taken into account, (ii) New correspondences which are not in the initial correspondence set are created during matching, (iii) It is assumed that template points i.e., points in the source image are sparsely distributed, and (iv) The same search range parameter is considered for all template point clusters. The subsequent sections show that accounting for these factors in image registration is central to non-affine document registration.



(a)



(b)

Figure 2.2 (a) Correspondences before RANSAC, (b) Correspondences after RANSAC. Wrong correspondence is shown in red color (cross marked). This outlier deviates the region of interest from the desired location.

2.2 Document Image Registration Methodology

Figure 2.3 presents an overview of the methodology that is designed to address some of the drawbacks described earlier. The rectangular boxes (blue color) in the template image of Figure 2.3 indicate the regions of interest to be extracted from a test image. Clusters of template points are formed using k-means [104] in the template image. Clusters that satisfy a proximity criterion with respect to the Regions Of Interest (ROI) are selected for registration. Furthermore, a histogram based uniformly transformed correspondence estimation is incorporated into the framework to speed up iterative correspondence estimation. The following subsections show how the prior knowledge of correspondences can be integrated into TPS-RPM framework, and enhance RANSAC and TPS-RPM by minimizing the registration error computed using local gradient information by demonstrating the performance of these algorithms for non-rigid deformations.

The rest of this section describes the methodology in detail. Sections 2.2.3 and 2.2.4 present the iterative approaches for outlier elimination and registration using RANSAC and enhanced RANSAC, respectively. Iterative approaches for non-rigid registration using TPS-RPM and enhanced TPS-RPM framework are presented in Sections 2.2.5 and 2.2.6, respectively.

2.2.1 Template Point Selection and Initial Correspondence

Extract invariant points from the template and test images using methods such as SIFT or SURF. Denote the points in template and test image by X and Y respectively. Cluster feature points in the template image using K-means algorithm [104]. For each ROI r in the template image, points belonging to m clusters that are closest to the ROI are selected as the template point set for the ROI (X_r). The idea behind the selection of points only in the m closest clusters is that these points move closely with the ROI, and further it reduces the non rigidity among the points.

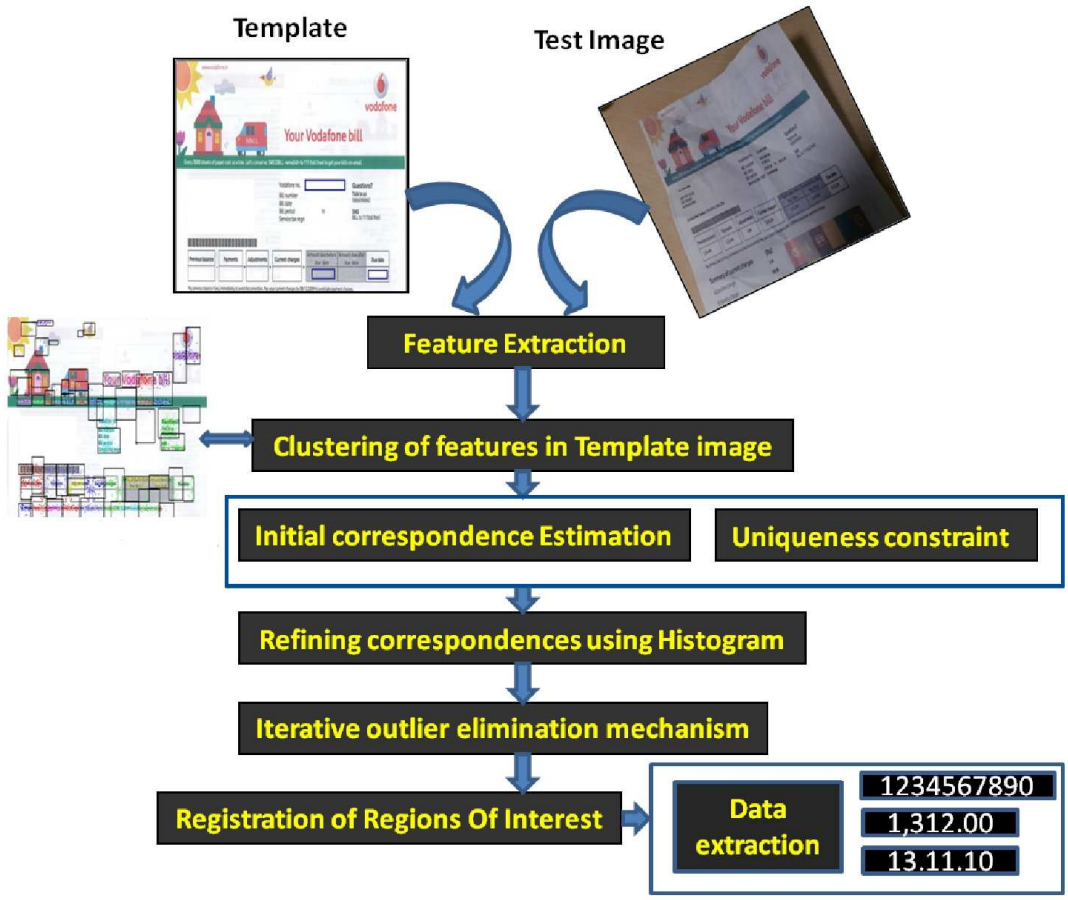


Figure 2.3 Overview of document image registration. The template image can be a scanned image or electronically generated where the Regions Of Interest (ROI) are known. Expected output is ROI in the test image.

Lowe's [28] method of initial correspondence generation is used to map points in X_r onto feature points in Y [28]. For each $x_i \in X_r$, two closest points in Y are found by using Euclidian distance in the feature space. If the ratio of these distances is less than t, then the template point with lesser distance is added to the correspondence set $C = \{(x_i, y_j) | x_i \in X_r \text{ and } y_j \in Y\}$. The correspondences now have a many-to-one mapping from X to Y. For each test point $y_j \in C$, a new correspondence set C' is obtained by performing a reverse mapping. Each point in $y_j \in C$ is now mapped onto the points $x_i \in C$. Correspondences are retained only if the obtained mapping is already present in C.

This ensures that for each $y_j \in Y$, there exists only one $x_i \in X_r$. The new correspondences are now $C' = \{(x_i, y_j) | x_i \in X_r, y_j \in Y$, and $(x_i, y_j) \in C\}$.

2.2.2 Refine Correspondence Set Using Histogram

Eliminate correspondences among outliers by using a histogram of Euclidean distances on the Cartesian coordinate space. The Euclidean distance between Cartesian coordinates of x_i and y_j for all $(x_i, y_j) \in C'$ is obtained and placed into histogram bins as shown in Figure 2.4. Bin size is given by $(max_{dist} - min_{dist})/$*(number of bins)*, where max_{dist}, min_{dist} are the maximum and minimum Euclidean distances of the corresponding points $(x_i, y_j) \in C'$ and *number of bins* is empirically set to ten. Correspondences whose euclidean distances fall in the peak bin and the bins that are within the threshold t_e (empirically set 80%) of the height of the peak bin are selected in a new set C''. This step operates under the assumption that while local distortions in document images can be non-planar, these distortions will not grossly alter the relative distribution of corresponding points. The results section will discuss how this step eliminates gross outliers, improving the convergence rate of iterative mechanisms. Figure 2.5(a) shows the one-to-one correspondences (C') obtained for a test image, Figure 2.5(b) shows refined correspondences (C''), and Figure 2.5(c) shows the correspondences after RANSAC under affine transformation.

2.2.3 Iterative Approaches for Outlier Elimination: RANSAC

RANSAC is an iterative optimization algorithm that repeats two phases: (i) generation of hypothesis by randomly sampling the data and (ii) hypothesis verification on data. Termination is done after a fixed number of iterations or when a termination condition is met [52]. Each RANSAC iteration selects three random non-collinear points from $x_i \in X$ such that $(x_i, y_j) \in C''$. Using the correspondence between x_i and y_j, an affine transformation matrix M is computed. The transformation matrix M is applied on

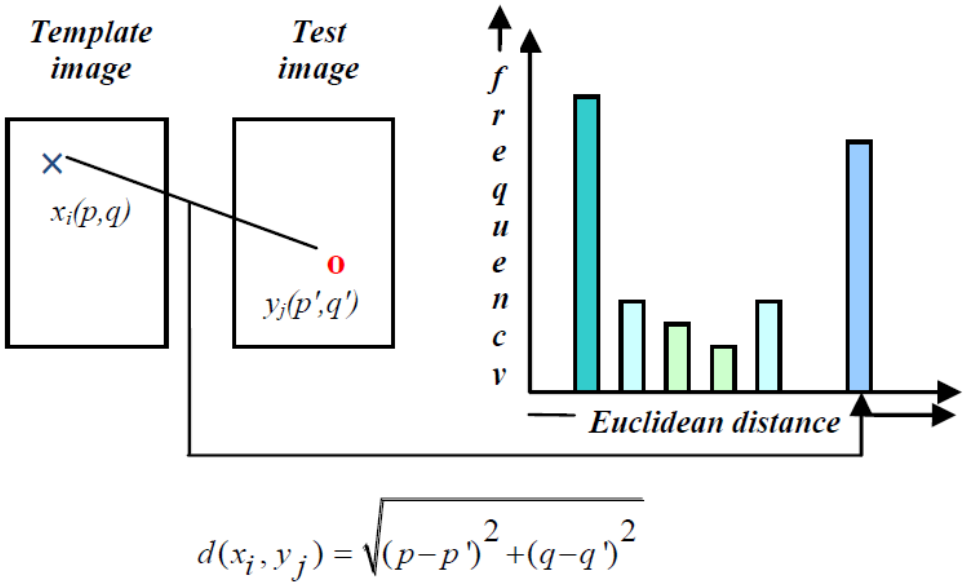
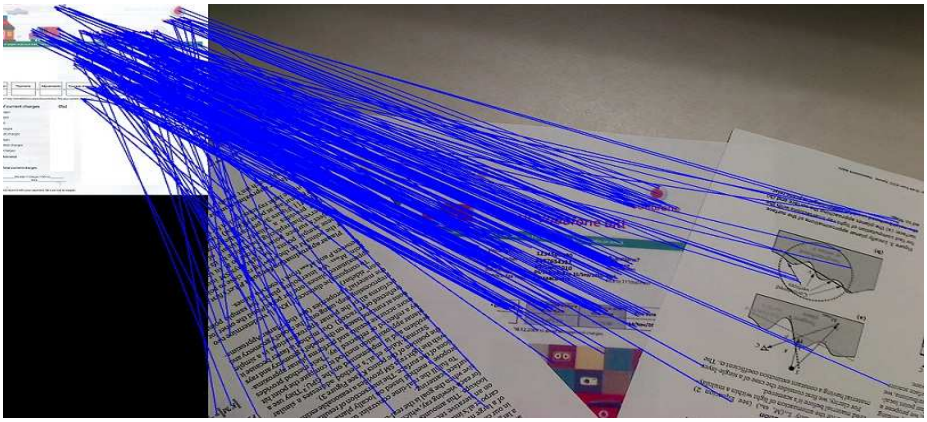


Figure 2.4 Correspondence estimation using Euclidean distance histogram.

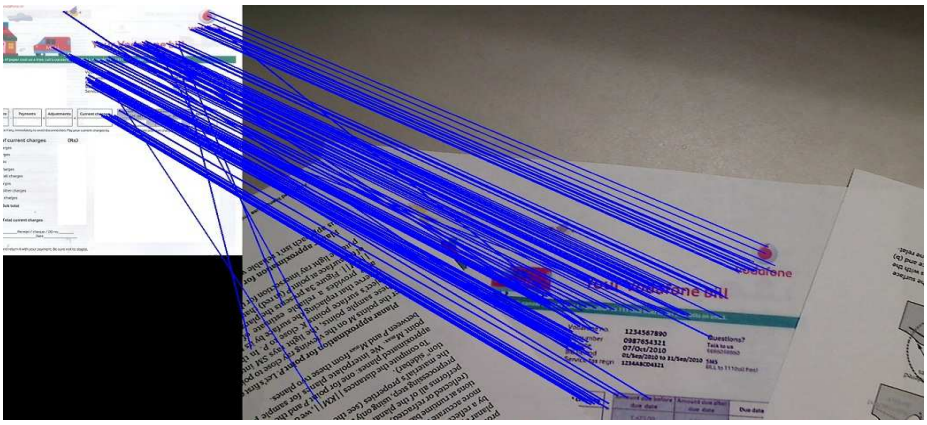
$\forall x_i | x_i, y_j \in C''$, to obtain $\bar{x_i}$. If $\bar{x_i} \equiv y_j$, x_i is marked as inlier, else x_i is marked as outlier. If the number of inliers in a particular iteration is greater than inliers in a previous iteration, the current set of inliers is accepted. The algorithm is terminated after a fixed number of iterations.

2.2.4 Enhanced RANSAC for Robust Registration

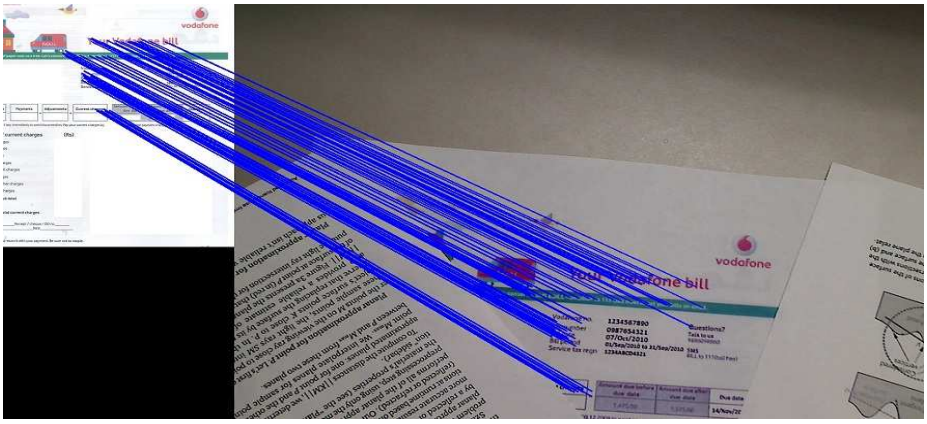
RANSAC is able to eliminate correspondences that do not conform to global geometry, and obtain a gross match between the template and test images. However, as mentioned earlier, an additional verification is needed to eliminate outliers arising from locally non-affine distortions as shown in Figure 2.2. Since specific regions of the image are of interest, processing is limited to ROI. In addition, assume that there will be image regions near the ROI that are similar across the test and template image. In each iteration of RANSAC, when the transformation matrix M is obtained, use M to warp the test image onto the template using cubic interpolation. Histogram of Oriented Gradients (HOG) [28] is computed from image regions surrounding the ROI in the template image and warped test



(a)



(b)



(c)

Figure 2.5 Correspondences at different stages of the framework (a) after Lowe's [28] method and one-one mapping, (b) after Euclidean distance based histogram, and (c) after RANSAC.

image. A modified RANSAC is performed using Chi-square [50] similarity of the HOG as the matching criterion. The method is described in Algorithm 1. Figures 2.6, 2.7, 2.8, and 2.9 illustrate the approach of Enhanced RANSAC. Figure 2.6 shows the original template image with marked ROI, clusters of SIFT points obtained using K-means, near by regions of ROI obtained from clusters of SIFT points, and a camera captured test image. Correspondences after the application of Euclidean distance histogram are shown in Figure 2.7. Figure 2.8 shows the warped images along with the extracted near by regions during intermediate iterations of enhanced RANSAC. Finally, Figure 2.9 shows the warped image obtained by enhanced RANSAC along with the projected ROI on the test image.

2.2.5 Thin Plate Spline-Robust Point Matching

While enhanced RANSAC is capable of addressing some of the deformations, methods like TPS-RPM have been specifically designed to derive non-rigid transformation functions [56, 94]. This section describes the TPS-RPM algorithm, and a few drawbacks of the method when applied to registration of ROI. Enhancements to TPS-RPM are provided in the subsequent section. Let $X = x_i : i = 1, 2, ..., N$ be a sparsely distributed template point set and $Y = y_j : j = 1, 2, ..., M$ be a relatively dense test point set. Both point sets are projected on a normalized Cartesian coordinate plane. TPS-RPM [56, 94] uses Gaussian mixture density to model the distribution of test points, while Gaussian cluster centers are determined by the template points. In order to robustly align the two sets, the algorithm performs deterministic annealing, where the temperature T of the annealing process acts as a search range parameter. At high temperatures the algorithm aligns the two point sets by preserving global structure of the template points. As T decreases, the search becomes local, where it accounts local deformations. It starts the annealing process with a larger T such that all the test points will be in the vicinity of template point clusters. At each T, it alternately estimates the correspondences and computes the underlying transformation function. It computes the probabilities of all test points being assigned to the template point

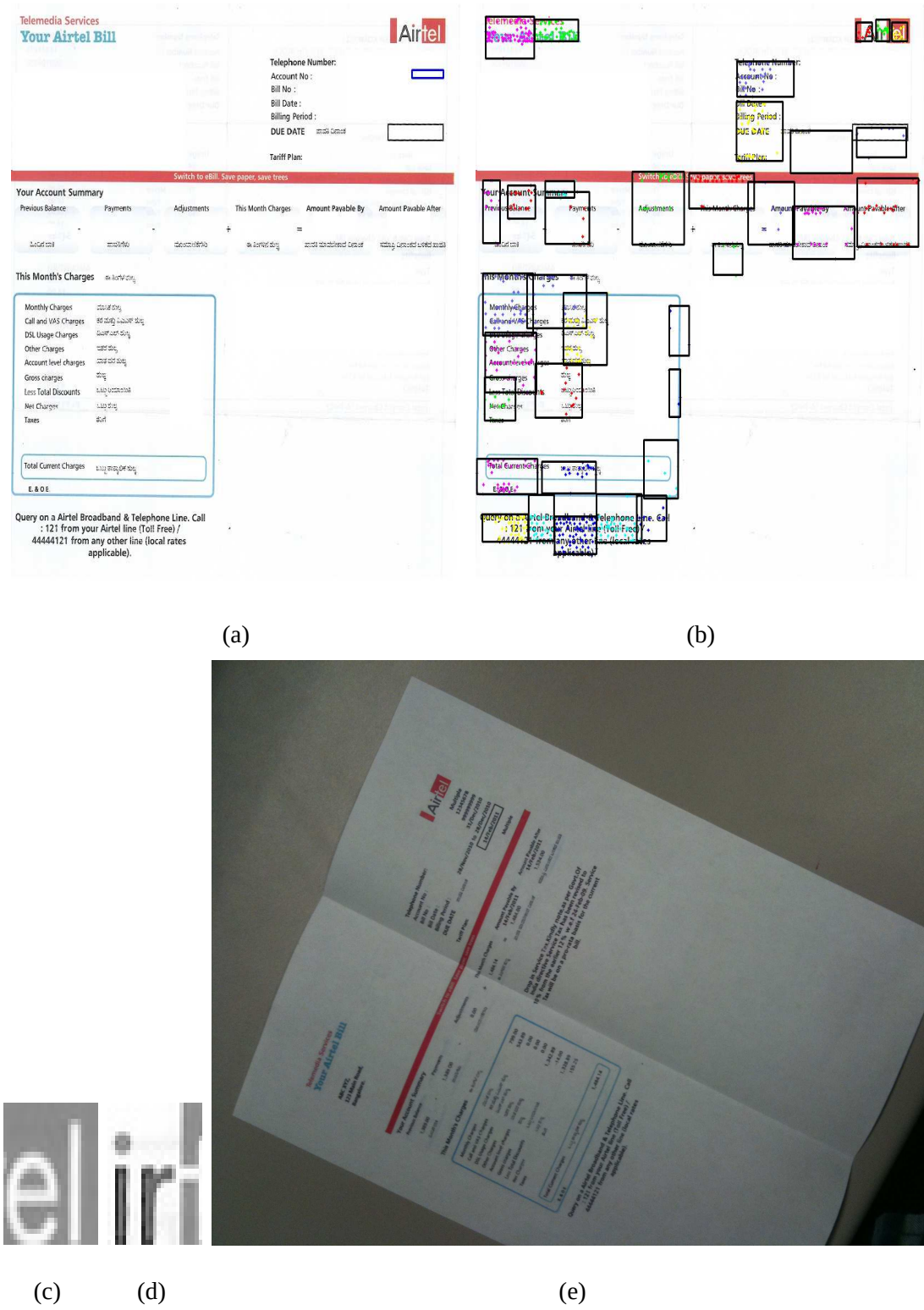


Figure 2.6 (a) Template image (ROI marked in blue rectangle), (b) clusters of SIFT points on (a), (c)-(d) near by regions for validation in enhanced RANSAC, and (d) test image.

Algorithm 1 Outlier Elimination Using Enhanced RANSAC

Input: Set of input correspondences C'', Test image B; m_r the number of fixed regions for the registration of ROI.

$HOG_i : i = 1, 2, ..., m_r$; HOG of fixed nearby regions.

HOG_{dist} : maximum positive integer.

Output: Refined correspondence set C''' with inliers, Transformation matrix M.

Initialization: $iterations$ = 0; $inliers$ = 0; $outliers$ = 0; MAX_{iter} = maximum number of iterations.

while $iterations < MAX_{iter}$ **do**

Hypothesis generation: Randomly select three correspondences among non-collinear points from C''. Compute the transformation matrix $Current_M$ from the three correspondences.

Hypothesis evaluation: Warp the test image B with $Current_M$ to align with the template image. Compute HOG of the fixed regions in the warped image HOG_j : $j = 1, 2, ..., m_r$

Compute the chi-square distance between HOG_i and HOG_j : $i, j = 1, 2, ..., m_r$, average it with m_r, and denote it as $Curr_{dist}$.

if $Curr_{dist} < HOG_{dist}$ **then**

Update:

$HOG_{dist} \leftarrow Curr_{dist}$

$M \leftarrow Current_M$

end if

end while

Update Correspondence set C''' with the correspondences that agree with M.

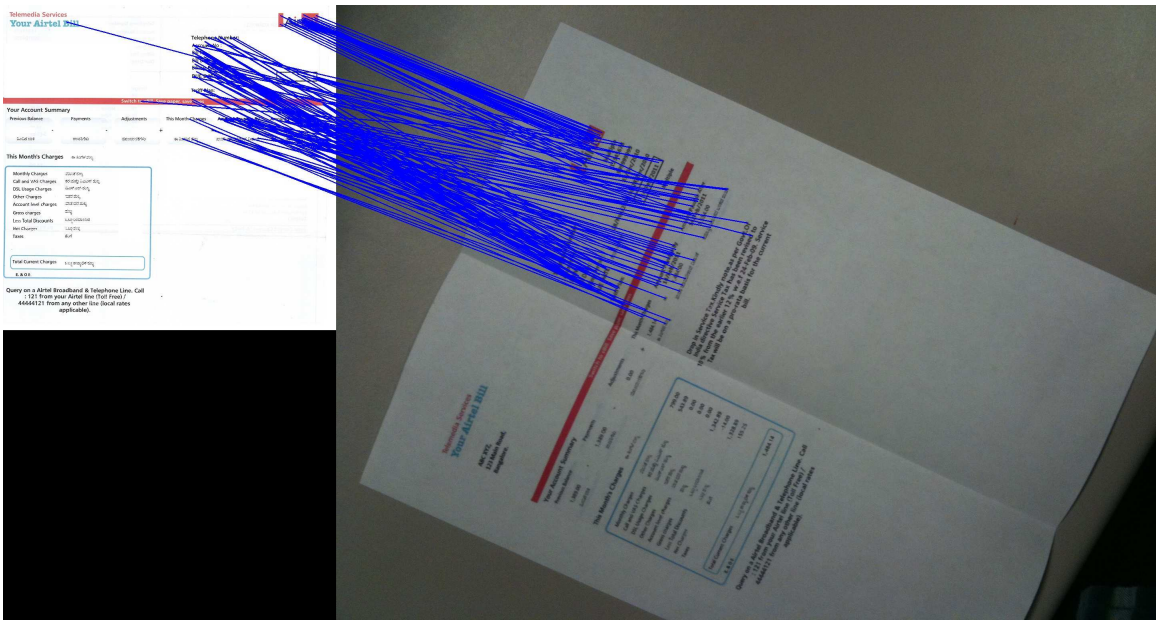
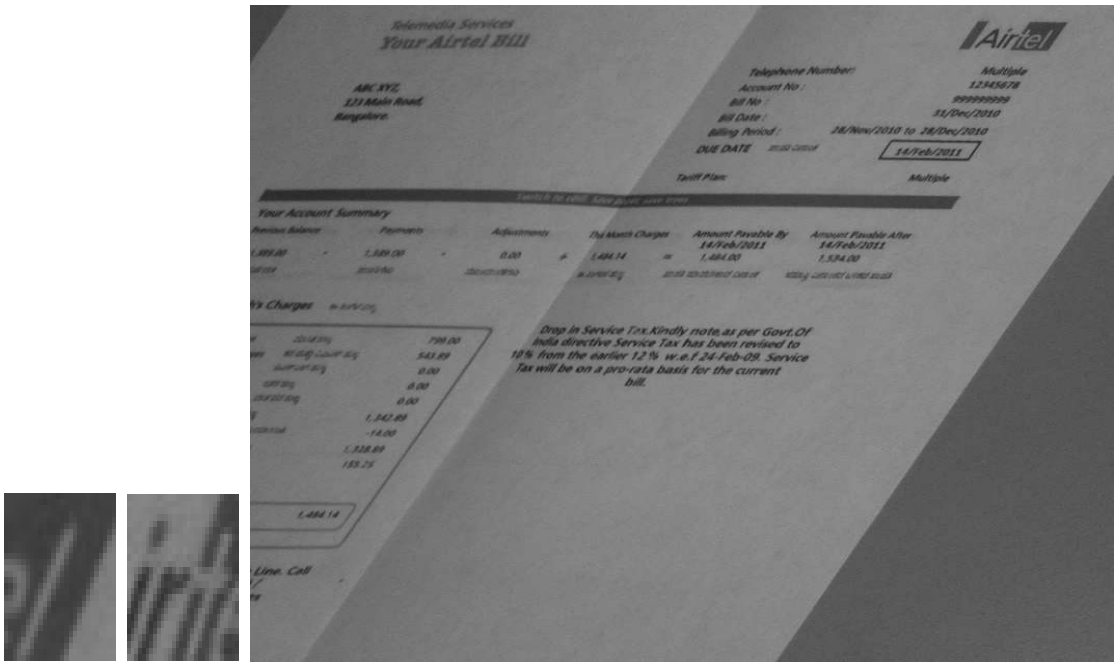


Figure 2.7 Correspondences after Euclidean distance based histogram while matching SIFT features extracted from Figure 2.6(d) with Figure 2.6(b).

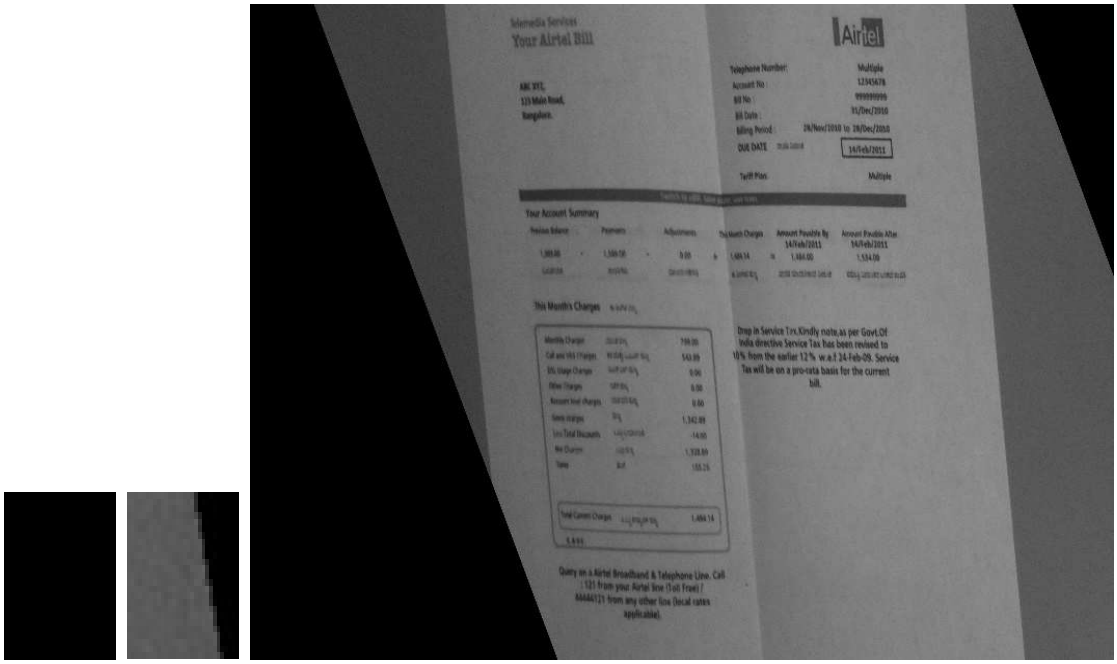
clusters and computes the probable location of the matching point for each template point. With the template points and the corresponding probable matching points, it estimates the transformation function f using TPS [59] to ensure smoothness in the transformation function. It repeats the annealing process with the template point clusters centered at $f(x_i)$ until T reaches final temperature T_{final} i.e., average of the squared distance between the nearest neighbors of the test points. To handle outliers in both point sets it maintains two additional clusters centered at the center of mass of the both point sets with large temperature T_0.

Drawbacks:

- The assumption of template point set as a sparsely distributed one is not true in the case of document images with multi-scale local features, as SIFT and SURF generates dense points in a given region.



(a) iteration 12



(b) iteration 28

Figure 2.8 Results of intermediate enhanced RANSAC iterations, extracted validation regions (two left columns) from warped images (right column) obtained by random sampling of correspondences.

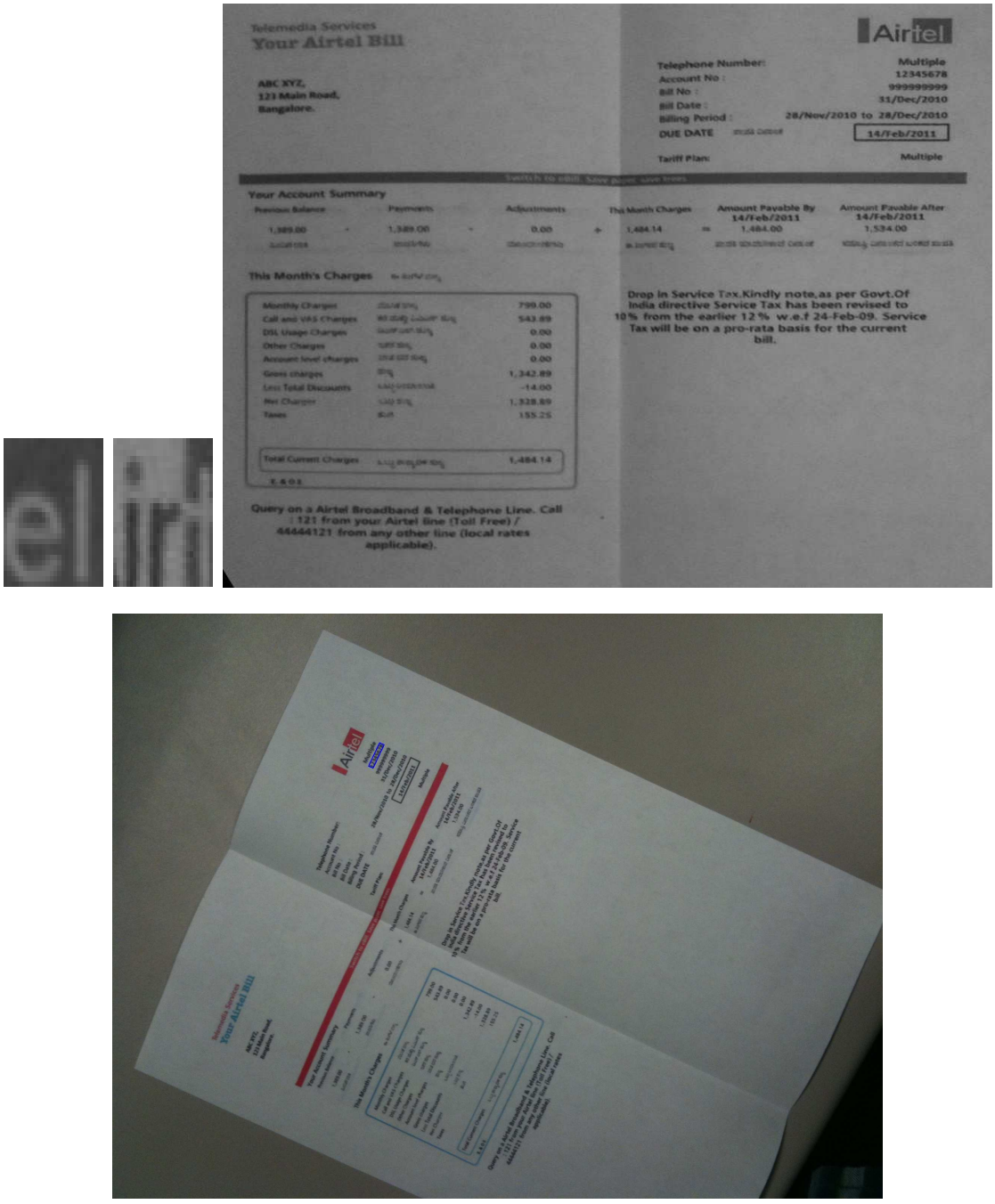


Figure 2.9 (a) Final warped image obtained by using enhanced RANSAC and extracted validation regions from it and (b) projected ROI on the test image using the transformation matrix obtained by the warped image in (a).

- TPS-RPM aligns the template point set to the test point set by considering only the geometry of the template point set. Apart from geometry there is an initial correspondence set which provides additional information to prevent template points being assigned to irrelevant test points.

- Each iteration of TPS-RPM generates new correspondences which are not in the initial correspondence set. The new irrelevant correspondences penalize the estimated transformation function.

2.2.6 Enhanced TPS-RPM

Enhanced TPS-RPM algorithm is designed to overcome the drawbacks of TPS-RPM. Apart from the template point set X_r and test point set Y, the algorithm takes into account the correspondence set C''. To prevent each template point being moved towards the irrelevant test point, it assigns different temperature T_i to each Gaussian cluster center x_i. Finally, the algorithm refines the new correspondences with nearby identical correspondences in C''. Remaining parts of this section present the problem formulation, enhanced TPS-RPM algorithm, and the refinement of new correspondences.

Let $C'' = (x_i, y_j)|x_i \in X_r, y_j \in Y$ be the set of input correspondences computed using the methodology in Section 2.2.2, where $X_r = x_i : i = 1, 2, ..., N$ and $Y = y_j : j = 1, 2, ..., M$ are the template and test point sets, respectively. As one-one mapping is enforced in the correspondence set, N is equal to M. Let f be the underlying Thin Plate Spline [59] based non-rigid transformation function, and the transformed template point set is $X'_r = x'_i = f(x_i) : i = 1, 2, ..., N$. Construct a correspondence matrix P to store the probabilities of each test point being assigned to each template point with dimension

$(N+1) \times (M+1)$.

$$P = \left(\begin{array}{ccc|c} p_{11} & \cdots & p_{1M} & p_{1,M+1} \\ . & \cdots & . & . \\ . & \cdots & . & . \\ . & \cdots & . & . \\ p_{N1} & \cdots & p_{1M} & p_{1,M+1} \\ \hline p_{N+1,1} & \cdots & p_{N+1,M} & 0 \end{array}\right) \tag{2.1}$$

The inner $N \times M$ sub-matrix defines the probabilities of each x_i being assigned to y_j. The presence of an extra row and column in the matrix handles outliers in both point sets. Each p_{ij} is computed as

$$p_{ij} = \frac{1}{T_i} e^{-\frac{(y_j - f(x_i))^T (y_j - f(x_i))}{2T_i}} \tag{2.2}$$

where $T_i : i = 1, 2, ..., N$ is the temperature of each template point cluster. For outlier clusters, the temperature T is kept at maximum throughout the annealing process. As discussed in Section 2.2.5, when T_i reaches T_{final} the correspondence is almost binary. If x_i is mapped to y_j then $p_{ij} \approx 1$. Similarly, if x_i is an outlier then $p_{i,M+1} \approx 1$, and if y_j is an outlier then $p_{N+1,j} \approx 1$. The matrix P satisfies the following row and column normalization conditions.

$$\begin{aligned} \textstyle\sum_{i=1}^{N+1} p_{ij} &= 1, \ for\ j = 1, 2, ..., M, \ and \\ \textstyle\sum_{j=1}^{M+1} p_{ij} &= 1, \ for\ i = 1, 2, ..., N \end{aligned} \tag{2.3}$$

The goal of the framework is to find an optimal transformation matrix P' and the optimal transformation function f' that minimizes the energy function $E(P, f)$ as defined below.

$$\begin{aligned}
[P', f'] &= \underset{P,f}{\text{argmin}}\ E(P, f), \\
E(P, f) &= E_g(P, f) + \lambda E_s(f) + E_a(P), \text{ where} \\
E_g(P, f) &= \textstyle\sum_{i=1}^{N} \sum_{j=1}^{M} p_{ij} ||y_j - f(x_i)||^2 \\
E_s(f) &= \textstyle\int\int \left[\left(\frac{\partial^2 f}{\partial u^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial u \partial v} \right)^2 + \left(\frac{\partial^2 f}{\partial v^2} \right)^2 \right] \\
E_a(P) &= T \textstyle\sum_{i=1}^{N} \sum_{j=1}^{M} p_{ij} \text{log} p_{ij} - \zeta \sum_{i=1}^{N} \sum_{j=1}^{M} p_{ij}
\end{aligned} \tag{2.4}$$

In the energy function E (Equation 2.4), $E_g(P, f)$ is the geometric feature-based energy term defined by Euclidean distance. $E_s(f)$ is the smoothness energy term with λ being the regularization parameter that controls smoothness of the transformation function. To favor rigid transformations at higher temperatures and local non-rigid transformation at lower temperatures, the framework reduces λ using an annealing schedule i.e., $\lambda_i = \lambda_{init} T_i$ where λ_{init} is a constant, $i = 1, 2, ..., N$. $E_a(P)$ is a combination of two terms; the first term controls fuzziness of P and the last term prevents too many points being rejected as outliers.

The transformation function f uses TPS [59], which can be decomposed into affine and non-affine subspaces, thereby accommodating both rigid and non-rigid transformations.

$$f(x_i, d, w) = x_i.d + \phi(x_i).w \tag{2.5}$$

where x_i is the homogeneous point representation of the 2D point x_i, d is a $(D+1) \times (D+1)$ affine transformation matrix of the D-dimensional image (For 2D images D=2), and w is a $N \times (D+1)$ warping coefficient matrix representing non-affine deformation. $\phi(x_i)$ is the TPS kernel of size $1 \times (N+1)$, where each entry $\phi_k(x_i) = ||x_k - x_i||^2 \text{log} ||x_k - x_i||$.

Algorithm 2 Enhanced TPS-RPM Pseudo Code

Input: Template point set X_r, Test point set Y, and the correspondence set C''.

Output: Correspondence matrix P and transformation $f = d, w$.

Initialize: Temperature $T_i : i = 1, 2, ..., N$ of each template point cluster with the Euclidean distance between the template point and the corresponding test point y_j specified in C'', T_{final} as average of the squared distance between the nearest neighbors of the test points.

Initialize: smoothness parameter $\lambda_i \leftarrow \lambda_0 T_i : i = 1, 2, ..., N$

Initialize: d with identity matrix, P using Equation. 2.2, and w with a zero matrix.

while $max(T_i) > T_{final}$ **do**

repeat

Update Correspondence: Compute P using Equation 2.2

Normalize P using Equation 2.3 iteratively.

Update transformation Update w and d using QR decomposition([55, 56])

until P, d and w converged

Update $T_i \leftarrow T_i \gamma$, update $\lambda_i \leftarrow \lambda_0 T_i; i = 1, 2, ..., N$; ($\gamma$ is the annealing rate)

end while

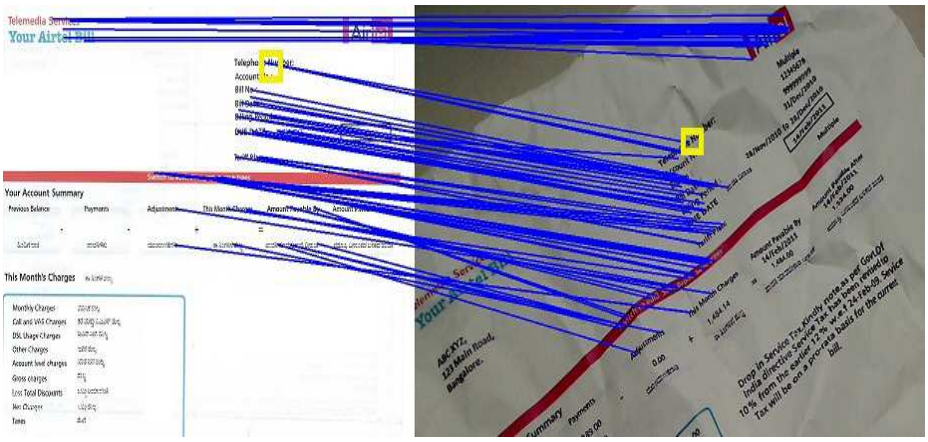
2.2.7 Refining New Correspondences

Even though the correspondence set is taken into account, the set of correspondences after the Algorithm 2 contains new correspondences which are not in C'', as the set C'' contains correspondences of the dense points. The new correspondences introduced by TPS-RPM lead to inaccurate transformation of the ROI i.e., blue boxes shown in Figure 2.10(b). To overcome this, refine the new correspondences with the correspondences of C'' that fall in the $h \times h$ window i.e., yellow boxes (h is empirically set to 15) of the new correspondence shown in Figure 2.10(a). Furthermore, refine the registration parameters obtained in Section 2.2.6 by minimizing the HOG error as described in enhanced RANSAC (Section 2.2.4). Figures 2.11(a) and 2.11(b) show the correspondences after enhanced TPS-RPM and the projected ROI, respectively.

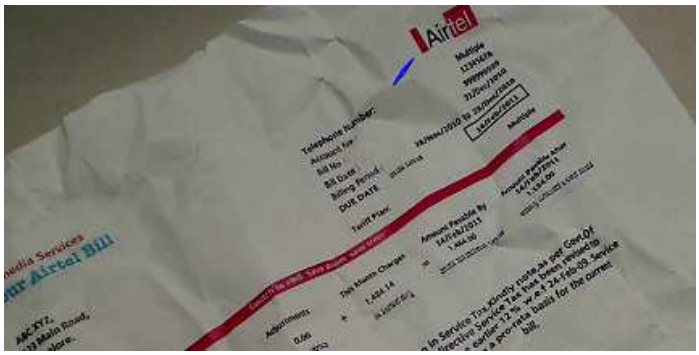
2.3 Results and Discussion

Experiments are conducted with twelve types of forms/utility bills falling in two different categories; one set is made of colored documents, shown in Figures 2.13 and 2.14, that have rich graphics and the other contains black and white documents, shown in Figure 2.12, that have minimal or no graphics. Test set consists of 480 images collected using two capturing devices iPhone3GS and Logitech webcam Pro 9000, 240 images using each capturing device. A few samples are shown in Figure 2.1. For each type of form, a template is collected from a color scanner at 150 dpi and locations of the required fields of interest are marked manually. During run-time, this template image and locations of the fields of interest are input to the registration algorithm. For each form, 20 test images are collected with each of the two capturing devices. The experiments use an AMD Athlon Dual core 2.69GHz machine with 1.75GB memory, taking on average 2 seconds to register each image excluding feature extraction.

Five approaches have been compared: RANSAC, RANSAC + Histogram, Enhanced RANSAC + Histogram, TPS-RPM, Enhanced TPS-RPM, and Enhanced TPS-RPM with



(a)



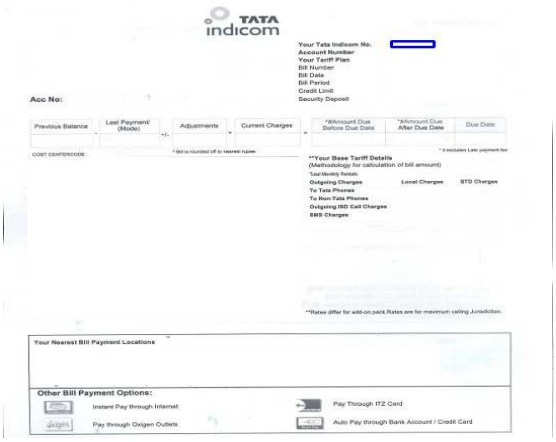
(b)

Figure 2.10 (a) Correspondences after enhanced TPS-RPM, (b) ROI from correspondences of (a) (ROI is in blue color at top right corner).

(a)

(b)

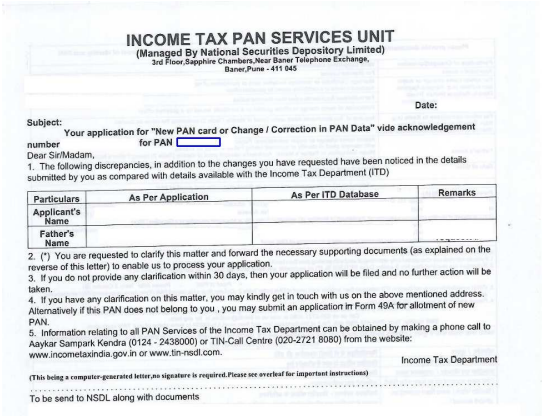
Figure 2.11 Refining correspondences using enhanced TPS-RPM. (a) correspondences after refinement (Section 2.2.7), and (b) ROI from correspondences of (a).



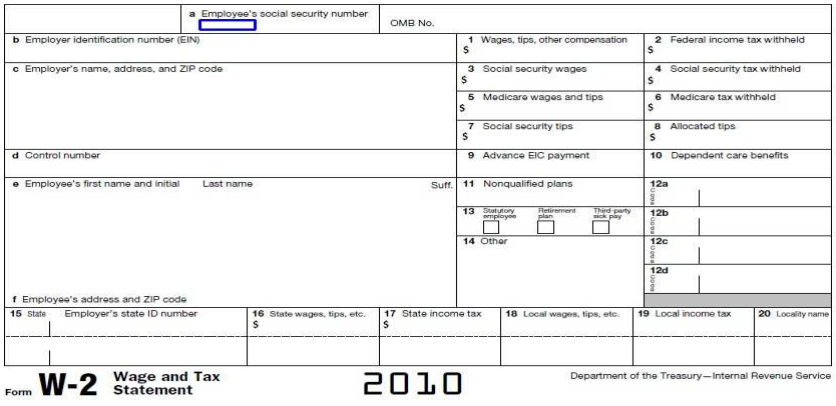
(a)

(b)

(c)



(d)



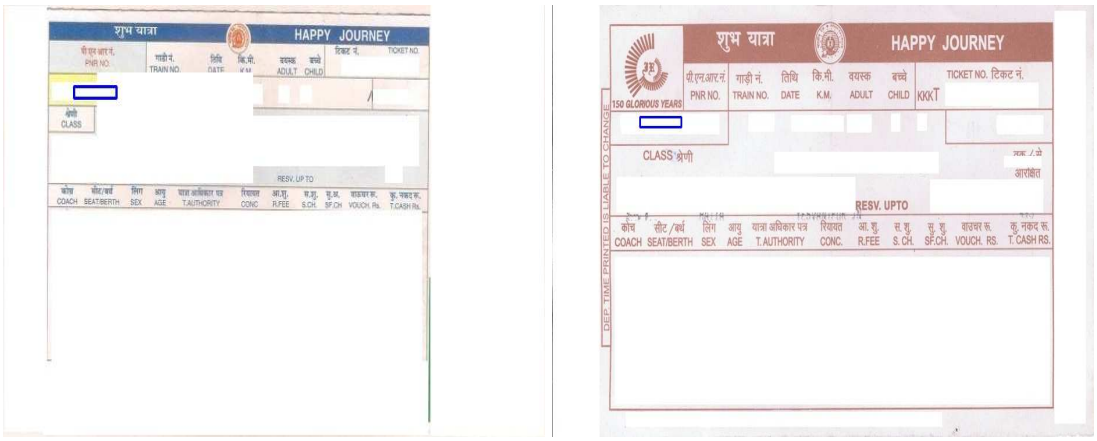
(e)

Figure 2.12 Black and white templates with minimal graphics along with ROI shown in blue rectangular boxes.

(a) (b)

(c) (d)

Figure 2.13 Color templates with graphics along with ROI shown in blue rectangular boxes.



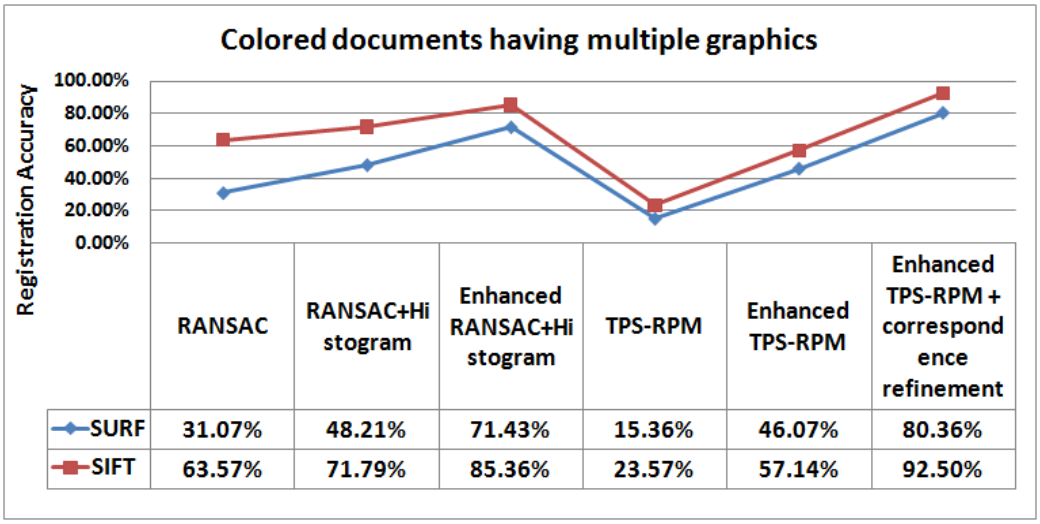
(a) (b)

(c)

Figure 2.14 A few more color templates with graphics along with ROI shown in blue rectangular boxes.

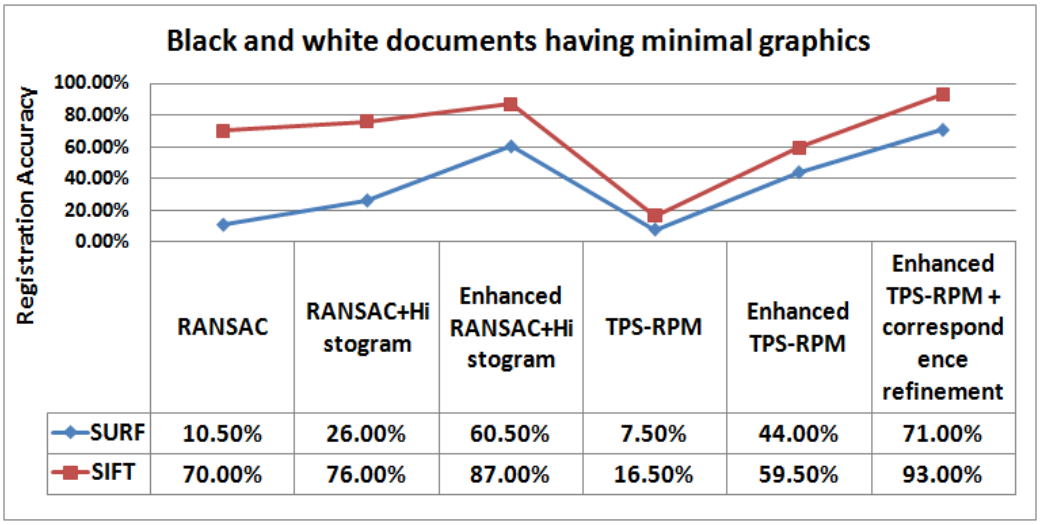
refinement of new correspondences. In RANSAC, algorithm presented in Section 2.2.3 is applied after obtaining initial correspondences (Section 2.2.1). In RANSAC + Histogram, RANSAC is applied after refining the initial correspondences using histogram of Euclidean distances (Section 2.2.2). In Enhanced RANSAC + Histogram, the enhanced RANSAC algorithm presented in Section 2.2.4 is applied after the Euclidean distance based histogram.

For RANSAC, RANSAC + Histogram, and Enhanced RANSAC + Histogram, threshold t of Lowe's approach is set to 0.9 and maximum RANSAC iterations is set to 100. In the case of the three methods based on TPS-RPM, Lowe's threshold is set to 0.6 and 0.8 for SIFT and SURF, respectively, to generate reasonably sparse points with enough correspondences. This difference in Lowe's threshold for SIFT and SURF comes from the fact that non-rigid registration depends on the selection of control points. Empirical evaluation shows that SIFT generates enough control points with small threshold value compared to SURF. Matching is restricted to points in the template image that fall in clusters close to an ROI. The set of template image points to be used for matching are selected in the following manner: (i) For each ROI, all clusters are marked as unselected, (ii) While the number of match points for the ROI is less than 300, the nearest unselected cluster is marked as selected and add points in this cluster to the set of match points for the ROI. Figure 2.15 shows the performance of different registration methodologies with SIFT and SURF features on different template images. Registration accuracy is measured as number of truly registered regions (90% overlap) divided by total number of regions. Enhanced RANSAC + Histogram and Enhanced TPS-RPM with refinement of correspondences outperforms the other methods. Enhanced TPS-RPM with refinement performs slightly better than Enhanced RANSAC + Histogram as it has the advantage of deriving complex transformation. TPS-RPM performance is poor, which is likely due to its assumption of sparseness in the point sets. In black and white images that are primarily white with sparse content e.g., W2 forms, SURF performs very poor on all the methods.



	RANSAC	RANSAC+Histogram	Enhanced RANSAC+Histogram	TPS-RPM	Enhanced TPS-RPM	Enhanced TPS-RPM + correspondence refinement
SURF	31.07%	48.21%	71.43%	15.36%	46.07%	80.36%
SIFT	63.57%	71.79%	85.36%	23.57%	57.14%	92.50%

(a)



	RANSAC	RANSAC+Histogram	Enhanced RANSAC+Histogram	TPS-RPM	Enhanced TPS-RPM	Enhanced TPS-RPM + correspondence refinement
SURF	10.50%	26.00%	60.50%	7.50%	44.00%	71.00%
SIFT	70.00%	76.00%	87.00%	16.50%	59.50%	93.00%

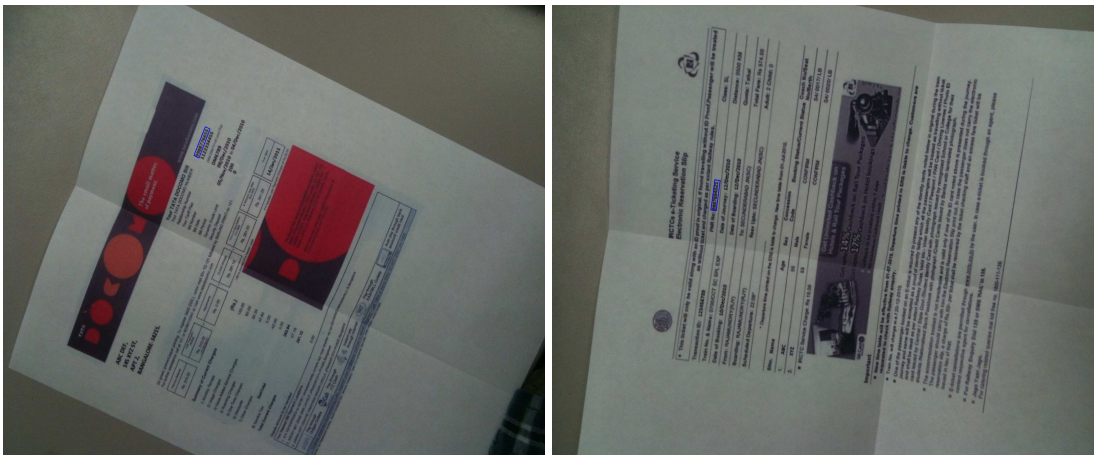
(b)

Figure 2.15 Comparison of registration methodologies using SIFT and SURF point features on different image types.

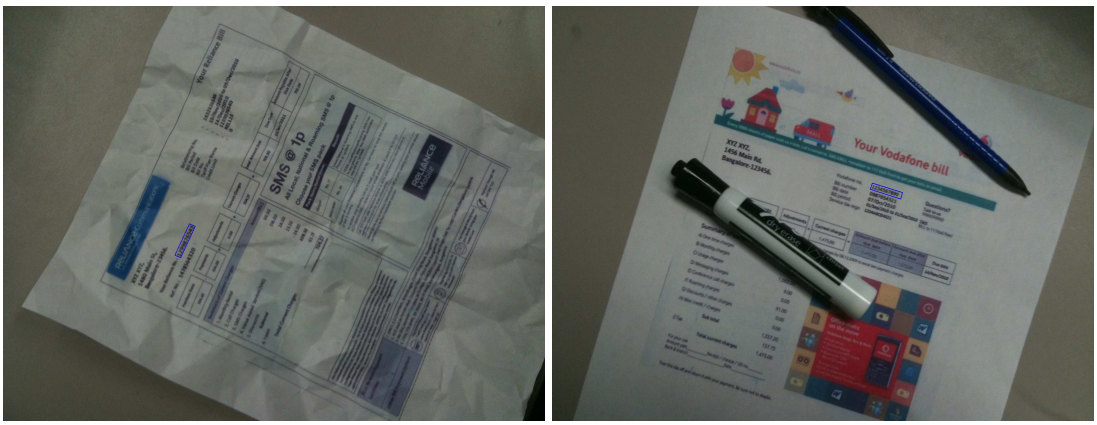
Euclidean distance Histogram as a preprocessing step to RANSAC significantly improves the performance of RANSAC on all the test cases. To test the effect of pre-processing steps on RANSAC convergence, RANSAC is terminated when 90% of the correspondences are inliers. Using Euclidean distance Histogram for pre-processing reduced the number of iterations by 60%, showing a positive effect on the convergence of RANSAC. Furthermore, SIFT gives larger number of control points surrounding the ROI with superior repeatability as compared to SURF. This is likely to be critical for non-rigid registration and leads to SIFT performing slightly better than SURF on all template types.

2.4 Conclusions

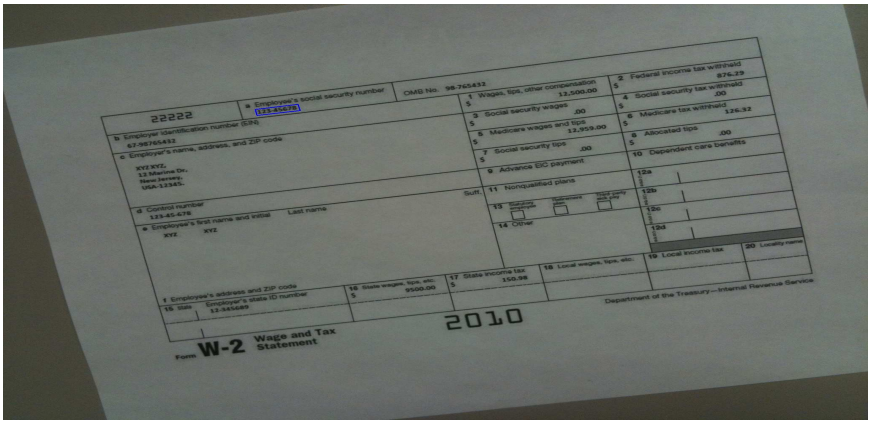
A framework for robust registration of camera captured document images is presented. Four novel aspects that comprise the framework are: clustering of feature points using K-means, Histogram based outlier refinement to speed up iterative algorithms, enhanced RANSAC for robust registration of document images, and finally enhanced TPS-RPM with refined correspondences for registration of images under non-rigid deformation. Clustering of feature points enables selection of nearby regions for registration of ROI. Euclidean distance based histogram not only eliminates the outliers but also enhances the convergence rate of RANSAC. Enhanced RANSAC algorithm refines the global registration parameters to suit each ROI, accommodating non-affine deformations. Enhanced TPS-RPM incorporates prior knowledge of correspondences into TPS-RPM and leads to better registration of non-rigidly deformed images. One limitation is that matching is applied to known ROI in the template image. While this is a reasonable assumption for several document processing applications, it is not a valid assumption in general.



(a) (b)



(c) (d)



(e)

Figure 2.16 Registered ROI in the images from the test set.

CHAPTER 3

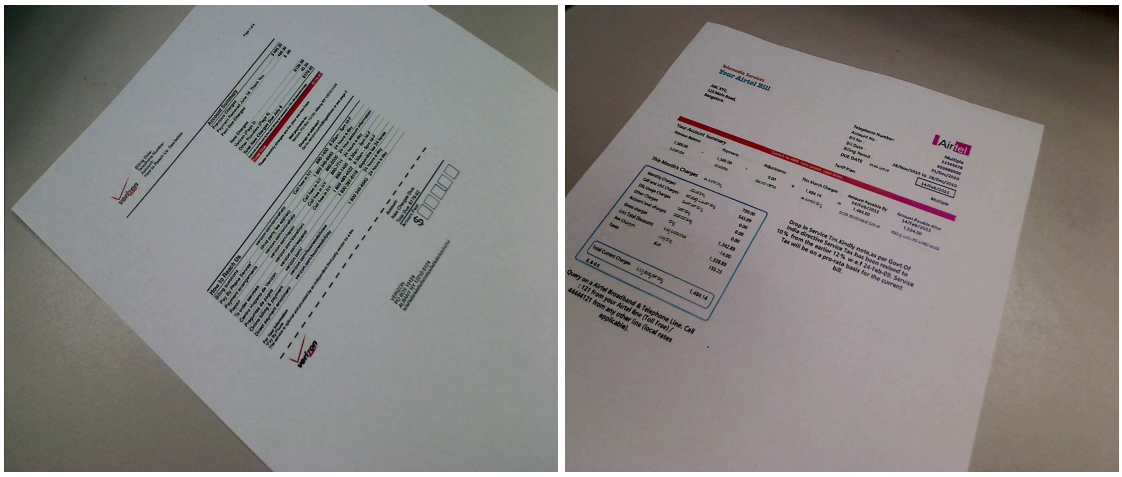
CATEGORIZATION OF CAMERA CAPTURED DOCUMENTS BY DETECTING LOGOS

This chapter presents a methodology to categorize camera captured documents into predefined logo classes. The existence of camera capturing noise such as intensity and large scale variations, partial occlusions, cluttering, and non-uniform folds make the detection task challenging. Besides, the appearance of logos is limited to a small portion of the captured document and a single document might contain more than one logo. The selection of robust local features and the corresponding parameters is presented by comparisons among SIFT, SURF, MSER, Hessian-Affine, and Harris-Affine. The evaluation of the methodology is conducted not only with respect to amount of space required to store the local features information but also with respect to categorization accuracy. Moreover, the methodology handles the detection of multiple logos on the document at the same time.

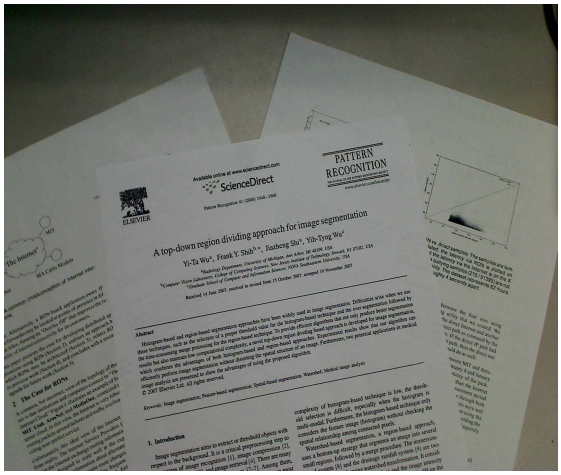
3.1 Related Work

Logos [19, 20, 22] play a vital role in uniquely identifying a document type. Generally, the appearance of logos is limited to a small portion of the document content and a single document might contain more than one logo as shown in Figure 3.1(b). Logo detection [19, 21, 22, 23] on scanned documents is a very well known problem in document analysis community. Most of these approaches rely on connected component extraction [19, 21, 22, 23]. A Bayesian approach by providing feedback between detection and recognition phases is specified in [22]. A method based on boundary extraction of feature rectangles to generate robust candidate logos is proposed in [21]. Geometric relationship among connected components is enforced in [19] to eliminate outliers. However, connected component extraction approaches rely on binarization [105, 106],

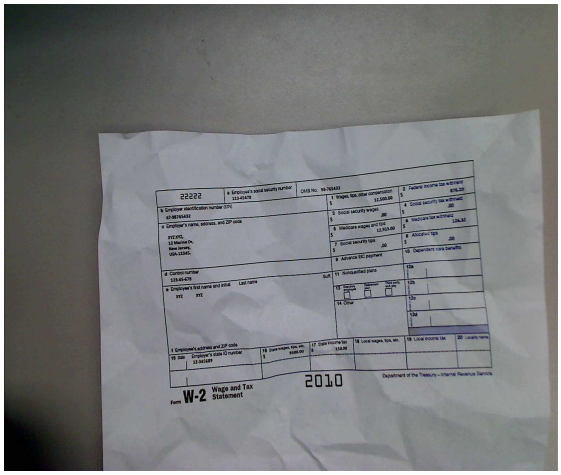
which introduces a lot of noise in camera captured documents. In [20], SIFT [28] features from a query image i.e., image under observation are matched against all the descriptors of logo classes. Though accuracies are reasonable on scanned documents, matching against all logo classes descriptors would not scale to large data sets, and is not a good strategy for real time applications.



(a) intensity and view-point variation



(b) background clutter and multiple logos



(c) crumples

Figure 3.1 Camera captured documents with logos.

On the other hand, literature in scalable document retrieval methodologies [17, 83, 10, 11, 12] rely on the entire document content including text, figures, and tables etc.

rather than just logos. A sequence of words is used as a query in [17]. A subregion of the document is used as a query in [10, 11, 12]. One common aspect of these approaches is that they need scanned documents as input. A document retrieval methodology using text is presented in [18]. Local invariant features are used to represent the predefined logo classes and the query document in order to overcome the challenges typically found in camera capture such as intensity variations, clutter, view-point variations, and crumples. Due to the availability of various local invariant features such as SIFT [28], SURF [25], MSER [30], Hessian-Affine [30], and Harris-Affine [30], there is always a question of selecting the robust feature.

The rest of the chapter is organized as follows: Section 3.2 presents the comparison of various local invariant features and the selection of one for the logo detection task. The detailed methodology of camera captured document categorization is presented in Section 3.3. Section 3.4 presents the experimental results on a challenging data set, which also discusses the impact of dimensionality reduction and representation of the features. Finally, Section 3.5 concludes the chapter.

3.2 Comparative Analysis of Local Invariant Features

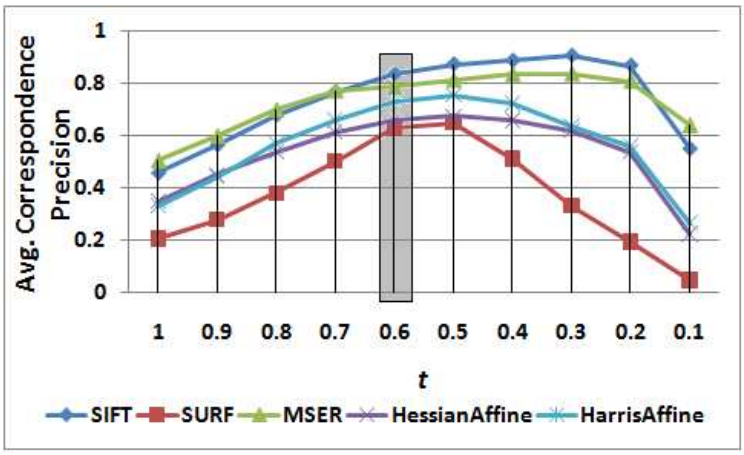
This section presents the selection of desired local feature by comparisons among various local invariant features. The features in consideration are SIFT [28], SURF [25], MSER [30], Hessian-Affine [30], and Harris-Affine [30]. The comparison is done using 25 logo classes and 125 camera captured documents with five documents under each logo class.

Let $L = \{L_1, L_2, ..., L_m\}$ be a set of logo classes, where m is the total number of logo classes. Each logo class L_i is represented by using n_i feature points $L_i = \{(x^j, y^j, f^j)\}$ for $j \in \{1, 2, ..., n_i\}$, where n_i is the total number of feature points in the i^{th} logo class; (x^j, y^j) and f^j are the Cartesian coordinates and d-dimensional description of the j^{th} feature point, respectively. Similarly, query image is represented as $Q = \{(x_q^j, y_q^j, f_q^j)\}$

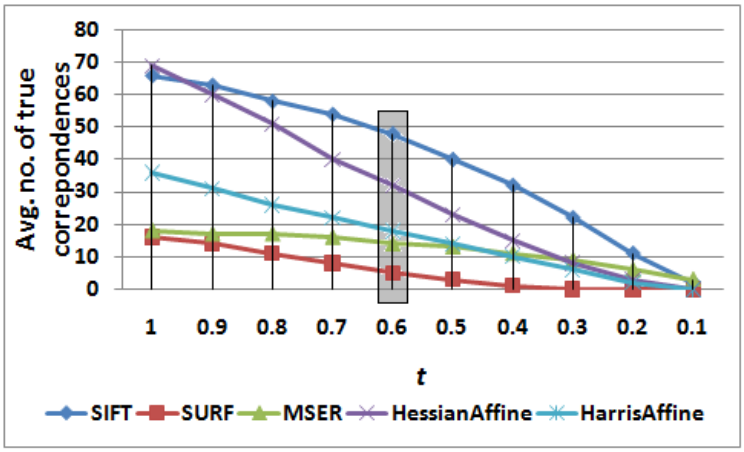
for $j \in \{1, 2, ..., n_q\}$, where n_q is the total number of features points extracted from the query document. Denote the j^{th} feature in L_i and Q as L_i^j and Q^j, respectively, and the corresponding d-dimensional feature descriptors as f_i^j and f_q^j, respectively. Lowe's [28] threshold t is used to make the comparisons, which is defined as the ratio of the distance between the logo descriptor and the first nearest neighbor among the query descriptors $f_q \in Q$ in the d-dimensional feature space to that of the second nearest neighbor.

$$t = \frac{D(f_i^j, f_q^{nn1})}{D(f_i^j, f_q^{nn2})} \tag{3.1}$$

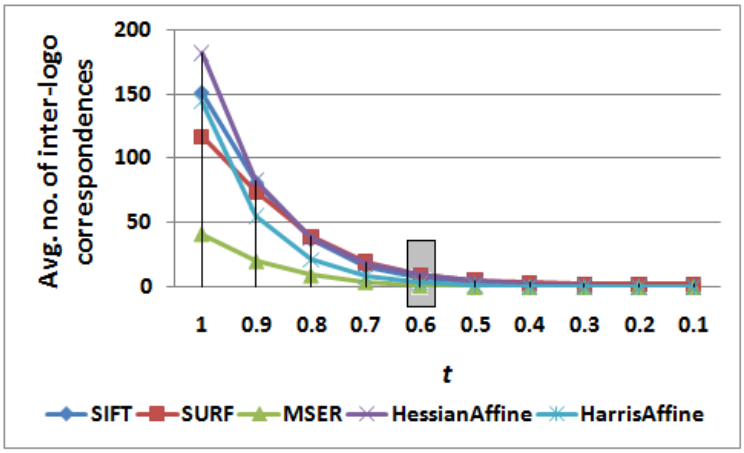
where $D()$ is the Euclidean distance in d-dimensional feature space, and $nn1$, $nn2 \in \{1, 2, ..., n_q\}$ are the indices of the first and second nearest neighbors to f_i^j in the feature space. A correspondence for each L_i^j is established with Q^{nn1} only if t is less than a predefined threshold, i.e., Q^{nn1} is the corresponding feature point to j^{th} feature of L_i in Q. As t goes down from 1 to 0, the ambiguity in the correspondences decreases, and more discriminative correspondences will be established. The behavior of the local invariant features is analyzed with respect to three important criteria: correspondence precision, number of true correspondences, and the number of inter-logo correspondences. Correspondence precision (as defined in Equation 3.2) and the number of true correspondences are analyzed by establishing the correspondences between each logo class and the corresponding five camera captured documents. The number of true correspondences is counted with the help of the established ground truth. Figures 3.2(a) and 3.2(b) show the behavior of average correspondence precision and average number of true correspondences at different thresholds t, respectively. A robust feature must have high average correspondence precision along with the large number of feature points to support partial occlusions and non-rigid deformations in the logo. Figure 3.2(c) shows the average number of inter-logo correspondences established with different feature types at various thresholds of t (for each logo class $L_i \in L$, the remaining classes $L_{i'} \in L; i \neq i'$ are used as queries). As some of the local features are common among multiple logos, using all the features will reduce



(a)



(b)



(c)

Figure 3.2 Comparisons among various local invariant features.

the discriminative power. One with lower number of average inter-logo correspondences should be preferred. From Figure 3.2, SIFT features at the shaded threshold t, i.e., 0.6, are the desired choice compared to the remaining features and thresholds. Section 3.3 presents an efficient logo-based categorization methodology using the derived feature type and the corresponding threshold t.

$$Correspondence\ Precision\ =\ \frac{Number\ of\ true\ correspondences}{Total\ no.\ of\ correspondences} \tag{3.2}$$

3.3 Methodology

The system has two modes of operation: off-line and on-line. Off-line mode is responsible for feature extraction from the logo classes, representation, and storage of the extracted data. On-line mode works in two stages. In stage 1, features are extracted from the query document and are matched against the features in the database to determine the candidate logo classes. In stage 2, top l candidate logo classes are then subjected to the cluster-based refinement process in the image space to eliminate false positives. Finally, the query document is categorized into the candidate logo classes left after stage 2. Figure 3.3 shows the overview of system configuration. The following subsections briefly explain the individual components of the system.

3.3.1 Off-line: Representation and Storage of Logo Class Features

Let $X = \{(x^j, y^j, f^j)\}$, $1 \leq j \leq n$ be the set of SIFT [28] features extracted from all the logo classes $L_i \in L$; where n is the total number of logo class features.

1. Dimensionality Reduction: This step is optional, and it reduces the dimensionality of SIFT [28] features. Generate a 128×128 dimensional matrix P with random numbers. Subject P to QR decomposition [34] to obtain the orthogonal matrix Q. The first r_d rows of the matrix Q form the projection matrix R. Project all the descriptors $f^j \in X$ onto R to reduce their dimensionality to r_d.

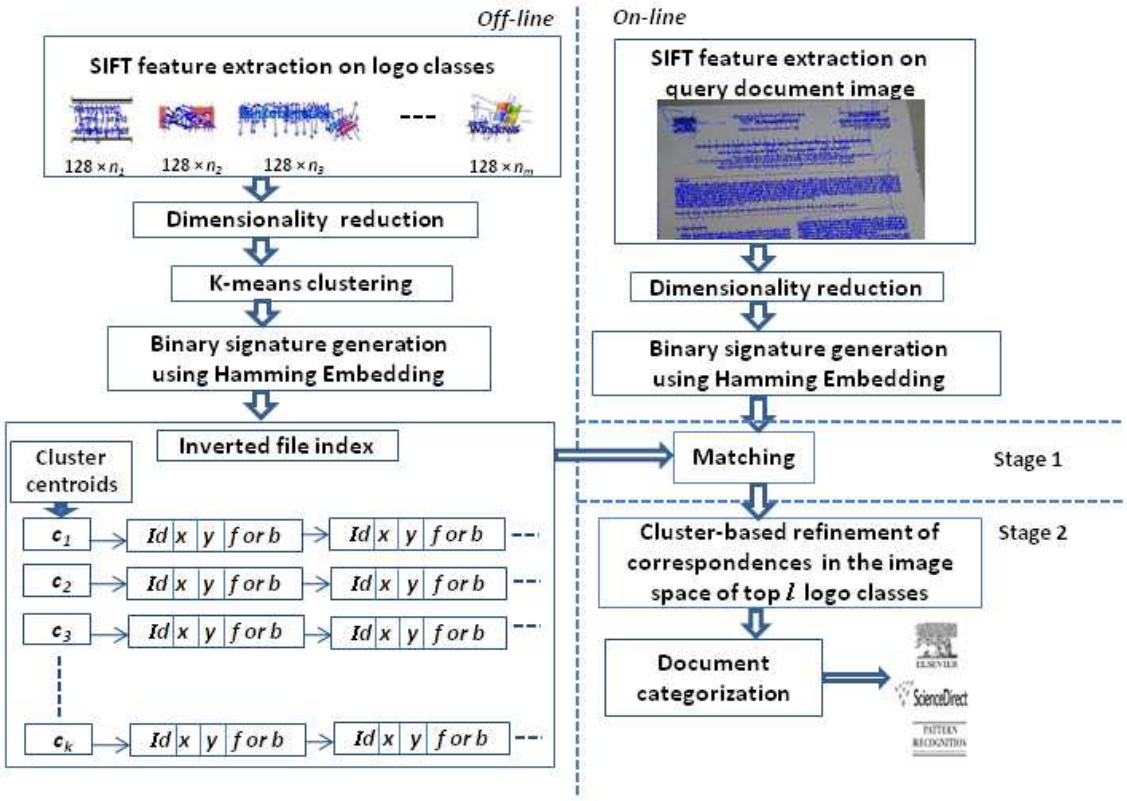


Figure 3.3 Document categorization framework.

2. Cluster Formation: Form the clusters of descriptors $f^j \in X$ in r_d-dimensional space using K-means [104], and denote the cluster centroids as $C = \{c_i | 1 \leq i \leq k\}$. The clusters are computed using SIFT features extracted from logo classes.

3. Hamming Embedding (HE): The main objective of this step is to convert the feature $f^j \in X$ into a binary string b^j for efficient representation, storage, and matching. For each r_d-dimensional descriptor $f^j \in C_i; 1 \leq i \leq k$, Hamming Embedding (HE) [34] is adapted to convert it to a bit string b^j of length r_d as defined in Equation 3.3.

$$\begin{aligned} b^j(x) &= 1, \quad if\ f^j(x) <= C_i(x);\ 1 \leq x \leq r_d \\ &= 0, \quad otherwise; \end{aligned} \tag{3.3}$$

4. Inverted File Indexing: Inverted file indexing [34, 63] structure is used to store the logo classes information. Only the cluster centroids $C_i \in C$ are indexed, and all the SIFT [28] features within each cluster are linked to their corresponding cluster centroid. The feature information attached is the logo class number(Id), Cartesian coordinates x^j, y^j, and the feature f^j (or) binary string b^j as shown in Figure 3.3. Denote the established index structure as I.

3.3.2 On-line: Feature Extraction on Query Document and Matching

Let $Q = \{(x_q^j, y_q^j, f_q^j)\}$, $1 \leq j \leq n_q$ be the set of SIFT [28] features extracted from the query document image and represented in the similar manner as logo class features (Section 3.3.1); where n_q is the total number of SIFT [28] features extracted from the query document. Algorithm 3 presents the mechanism of matching features in Q with the established inverted file index I of Section 3.3.1.

Refinement of Scores using Neighborhood Check: As the scores after stage 1 matching contain lot of outliers, refine the established correspondences in the top l candidate logo classes using cluster-based neighborhood check in the image space. Figure 3.4 shows matches established during Stage 1 matching. One can enforce the ordering among the

Algorithm 3 Stage 1 Matching

Input: Inverted File Index I(Section 3.3.1), Query features Q.

Output: Scores $S_i \in S; 1 \leq i \leq m$ of the logo classes.

Initialize: All $S_i \in S$ to zero.

for all $Q^j \in Q$ **do**

Determine the nearest cluster $C_i \in I$;

Initialize: D (Distance to all features $\in C_i$) to zero.

for all $(b^z|f^z) \in C_i$ **do**

Compute the distance D^z = D($b^z|f^z$, Q^j); where D() is $xor()$ for b^z, and Euclidean distance in r_d-dimensional space for f^z;

end for

sort D in decreasing order;

Increment the score $S_{Id(D^1)}$ by 1 only if $(D^1/D^2) \leq t$; where D^1 and D^2 are the distances to the first and second nearest features of Q^j, and t is Lowe's [28] threshold;

end for

sort S in decreasing order;

local features [35], and check for the relative order consistency between query document and the candidate logo class, or refine the correspondences by fitting a transformation model [28] to the correspondences. Due to the non-rigid deformations i.e., crumples, a cluster-based neighborhood check is applied in the image space to determine the outliers. Algorithm 4 presents the underlying mechanism. Figure 3.5 shows the matches refined after neighborhood check.

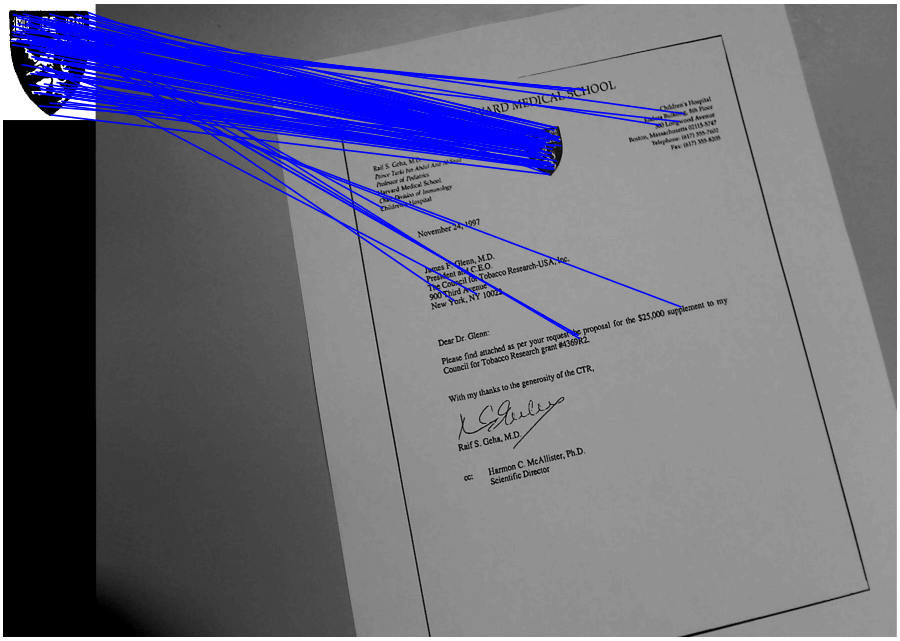


Figure 3.4 Matches established during Stage 1 matching.

3.4 Experimental Results and Discussion

Test set consists of 375 camera captured query documents of resolution 1600×1200 belonging to 25 logo classes. Figure 3.6 shows the logo classes and their distribution in the test set. F–measure [64] as defined in Equation 3.4 is used to evaluate the methodology. F–measure combines both recall and precision into a single measure, which is well

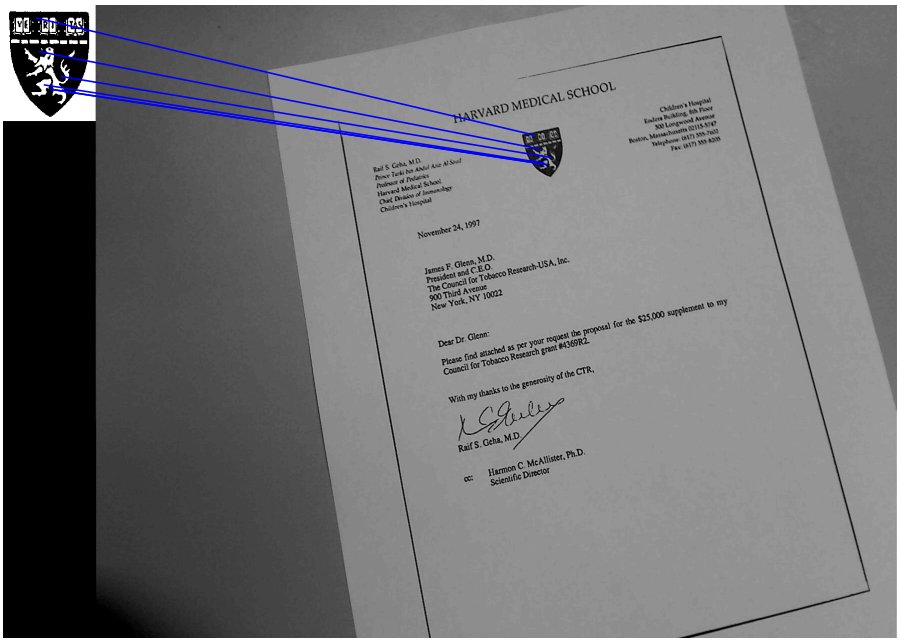


Figure 3.5 Matches established after neighborhood check.

Algorithm 4 Stage 2 Matching: Cluster-Based Neighborhood Check

Input: Top l candidate logo classes $L' \in L$ after stage 1 matching, and the corresponding scores $S' \in S$.

Output: Refined Scores S' of the candidate logo classes.

for all $L_i \in L'$ **do**

Initialize: neighborhood cardinality r_e to $\lceil$ sqrt(S'_i)$\rceil$.

repeat

for all features $(x^j, y^j) \in L'_i$ **do**

Let $N(x^j, y^j)$ and $N_q(x^j_q, y^j_q)$ be the r_e neighborhood features of the j^{th} correspondence between the logo class L'_i and the query document Q respectively;

Determine the probability of j^{th} correspondence being an inlier as $P^j = \frac{(N(x^j,y^j) \cap N_q(x^j_q,y^j_q))}{r_e}$;

Mark the j^{th} correspondence as inlier if $P^j \geq t_p$; where threshold t_p is set to 0.5;

end for

Update the correspondences in L_i with inliers, and refine the S'_i with the cardinality of L'_i i.e., $\|L'_i\|$;

until $S'_i \leq 3$

end for

sort S' in decreasing order, and eliminate all the logo classes $L'_i \in L'$ with the scores $S'_i \leq 3$;

informative compared to individual recall and precision scores. The higher the F–measure, the better the categorization accuracy. Figures 3.8 and 3.9 show the established matches of the images from the dataset with the corresponding logos.

$$\begin{aligned} Recall &= \frac{Number\ of\ true\ categories\ retrieved}{Total\ number\ of\ true\ categories} \\ Precision &= \frac{Number\ of\ true\ categories\ retrieved}{Total\ number\ of\ identified\ categories} \\ F\text{–}measure &= 2 \times \frac{Precision \times Recall}{Precision + Recall} \end{aligned} \tag{3.4}$$

Table 3.1 shows the accuracies at different stages, and different SIFT [28] feature

(a) 15 documents

(b) 15 documents

(c) 15 documents

(d) 15 documents (e) 135 documents (f) 45 documents (g) 105 documents

Figure 3.6 Logo classes and their distribution in test set.

representations with k = 100, t = 0.5, and l = 5. HE-128 and HE-64 in the Table 3.1 corresponds to feature representation with Hamming Embedding (HE) and bit string lengths of 128 and 64, respectively. From the Table 3.1, as the dimension of the SIFT [28]

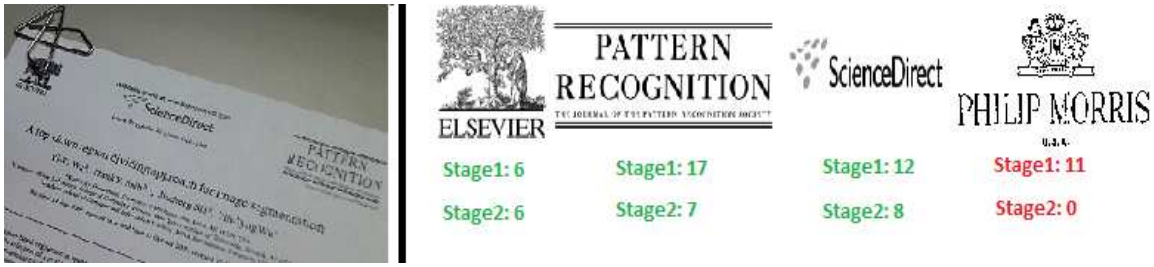


Figure 3.7 Category identification: left:query document, right: predicted categories (true: scores in green, false: scores in red).

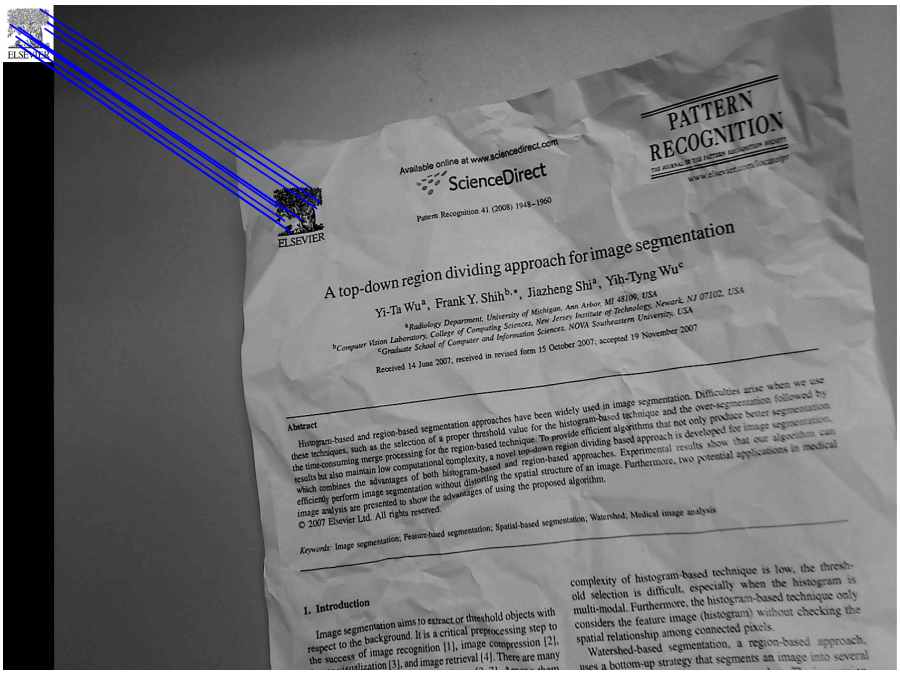


Figure 3.8 Matches established for Elsevier logo.

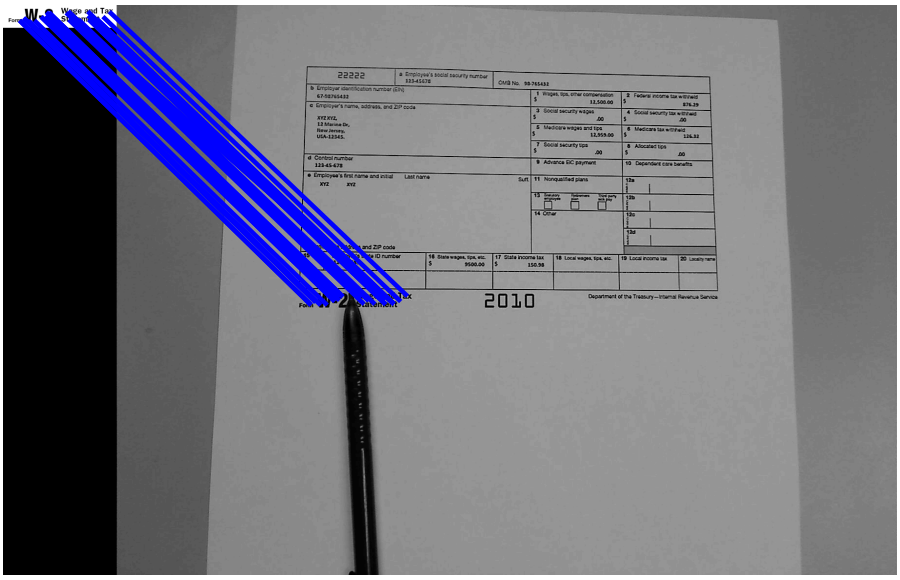


Figure 3.9 Matches established for W2 logo.

features decreases from 128 to 16, the corresponding stage 2 F–measure decreases gradually, and stage 2 matching significantly improves the stage 1 matching F–measures. HE with 128-bit string representation achieves a reasonable F–measure accuracy of 68.24% with enormous savings in storage. A similar kind of pattern is observed at k = 50 and k = 200, with a minor change of 1 to 2% in the F–measure, and slightly higher accuracies with increasing number of clusters k. The derived threshold t = 0.6 is also empirically verified by a comparison among other threshold values, and observed higher F–measure accuracies at t = 0.6. A F–measure accuracy of 36.54% is achieved by directly adapting the HE method of [34] with 128 bits and the specified parameters. Furthermore, the methodology is verified on Tobacco-800 [107] dataset and achieved a 95.14% F–measure accuracy as opposed to 92.5% using [19]. Finally, Figure 3.7 shows the scores of the identified categories of a query document at each stage. On an average, it takes 1 second to categorize the given query document on Intel core 2 duo machine using MATLAB.

Table 3.1 Accuracies at Different Stages of Matching and Different Feature Representations

	Feature Representation (dimensions)					
	16	32	64	128	HE-64	HE-128
Average Recall						
Stage 1	72.31%	84.18%	85.69%	88%	70%	83.24%
Stage 2	63.24%	76.13%	79.2%	81.07%	62.58%	78.13%
Average Precision						
Stage 1	28.35%	44.48%	40.49%	50.32%	21.17%	30.77%
Stage 2	48.94%	69.24%	69.3%	75.07%	57.55%	60.57%
Average F–measure						
Stage 1	40.73%	58.21%	54.99%	64.03%	32.51%	44.93%
Stage 2	55.18%	72.52%	73.92%	77.95%	59.96%	68.24%

3.5 Conclusions

A methodology to categorize camera captured documents based on logo detection is presented. The selection of robust features is done by comparisons among various local invariant features. The methodology not only categorizes the captured document under partial occlusions, intensity variations, and non-rigid deformations but also identifies multiple categories if present. Evaluation of methodology is presented with respect to different feature representations.

CHAPTER 4

SEGMENT-WISE MATCHING FOR CATEGORIZATION

This chapter presents a segment-wise matching approach for categorization of camera captured documents into predefined logo classes. SIFT is used to represent logo classes and query document in order to overcome the challenges typically found in camera capture such as intensity variations, clutter, view-point variations, and crumples. To obtain higher recall and precision accuracies, segmentation of query document image is presented by grouping area under intersecting dense affine covariant regions to maximize the margin between the matching scores of true logo classes and the rest. Besides, multiple descriptions of each feature that belong to different dominant orientations in the surrounding region are grouped and Hamming Embedding (HE) is applied to suppress the noise during descriptor quantization. Experimental results on a challenging dataset demonstrate a peak 13.25% increase in the F–measure accuracy compared to the methodology presented in previous chapter.

4.1 Motivation

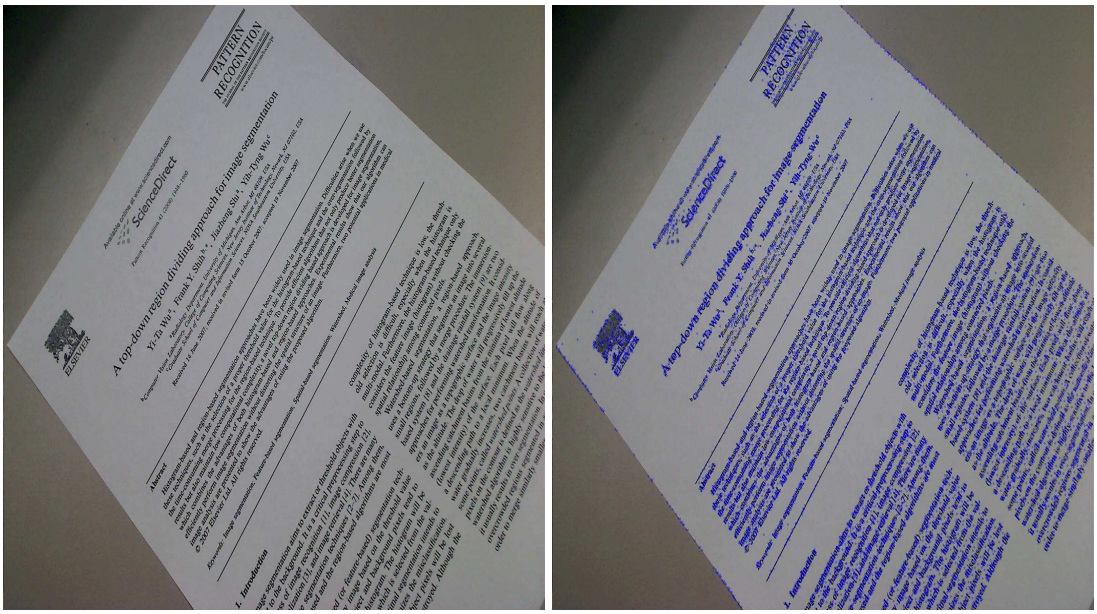
SIFT features are empirically shown robust to a wide variety of challenges such as background clutter, intensity variations, view-point variations, and crumples in Chapter 3. The methodology categorizes query document into predefined logo classes in a two stage matching fashion. In the first stage, local features from the entire query document are matched to determine candidate logo classes. Neighborhood check of computed matches is performed in second stage to refine retrieved candidate logo classes. Generally, the cost of performing second stage matching, which typically accommodates outlier elimination mechanisms, increases with increase in the number of false matches. Most of these false matches arise from using feature matches from the entire query document. Figure 4.1

illustrates the motivation to conduct matching limited to logo regions. An example query image and the corresponding SIFT features extracted from it are shown in Figure 4.1(a) and Figure 4.1(b) respectively. Figure 4.1(d) shows the partitioning of the number of matches of Figure 4.1(b) with Figure 4.1(c) to different regions of the query document. From Figure 4.1(d), it is clear that the corresponding logo region i.e., Elsevier accommodates more number of matches compared to other regions. Limiting the matches to those that arise from true logo region not only helps to increase the performance of outlier elimination techniques but also gives an approximate position of the logo class in the query document. Similarly, Figure 4.2(b) shows that the region containing pattern recognition logo contains more number of matches compared to other regions when matching Figure 4.1(b) with Figure 4.2(a). Furthermore, by distributing the matches to different regions of the query document and selecting a region with more number of matches reduces the number of matches of an irrelevant logo class. In this chapter, an efficient methodology to categorize camera captured documents into predefined logo classes is presented by limiting the matching to segments achieved by grouping area under intersecting dense affine covariant regions [30].

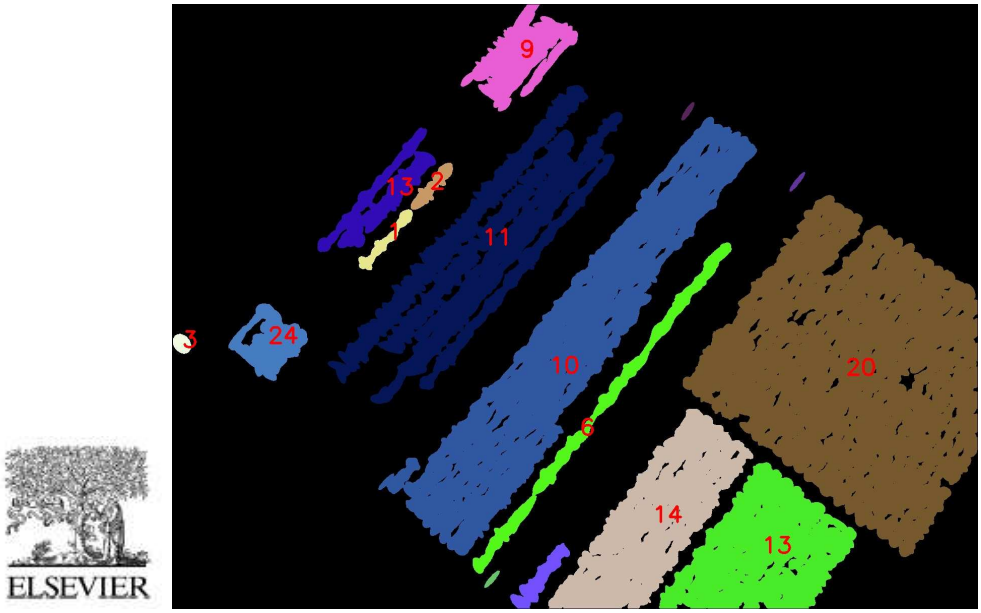
The rest of the chapter is organized as follows: Section 4.2 presents feature extraction and grouping of descriptors belonging to same feature. Inverted index computation of logo classes is presented in Section 4.3. Section 4.4 presents detailed methodology of camera captured document categorization. Section 4.5 presents experimental results on a dataset of real camera captured documents. Finally, Section 4.6 concludes the chapter.

4.2 Feature Extraction and Grouping

SIFT is used to represent logo classes and query document. SIFT chooses interest points at the extremum of difference-of-Gaussian scale space [42, 28] and describes the region around the interest points invariant to rotation. Given an image A, let $X = \{(x^j, y^j, f^j)\}$, $1 \leq j \leq m$ be the set of SIFT feature descriptors extracted from A; where m is the

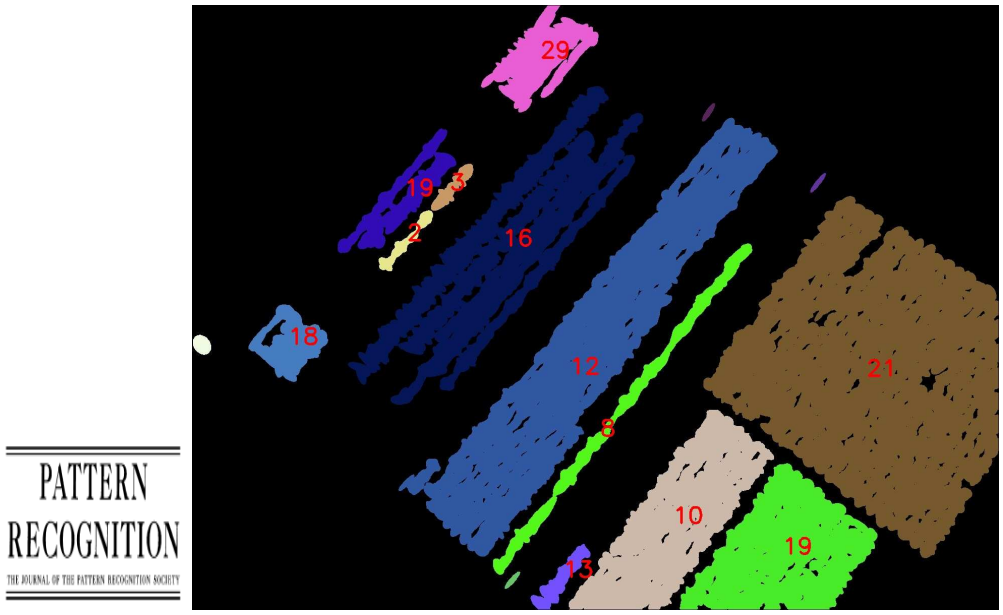


(a) (b)



(c) (d)

Figure 4.1 (a) Query image, (b) SIFT features extracted from (a), (c) Elsevier logo, (d) matched features of (b) with (c).



(a) (b)

Figure 4.2 (a) Pattern Recognition logo and (b) matched features of Figure 4.1(b) with (a).

total number of descriptors extracted, (x^j, y^j) denotes feature point position in A, and f^j represents corresponding d-dimensional description. SIFT describes surrounding region around interest points with respect to all dominant orientations in that region [28]. This leads to a state where each feature point has one or more feature descriptions associated with it. Figure 4.3 shows SIFT features extracted from example logo classes, one can observe that some feature points have multiple dominant directions (arrows in different directions at the same feature point). While matching a set of SIFT features extracted from one image with another set of SIFT features extracted from second image, these isolated descriptors could lead to false matches. To suppress such kind of false matches, descriptors corresponding to same feature point and supporting region i.e., scale are grouped. Refine X such that $X = \{(x^j, y^j, \{f^j\})\}$, $1 \leq j \leq r$; where r is the total number of feature points extracted, and $\{f^j\}$ is the set of all d-dimensional descriptions corresponding to feature point (x^j, y^j). The rest of the chapter uses X^j to denote j^{th} feature in X.

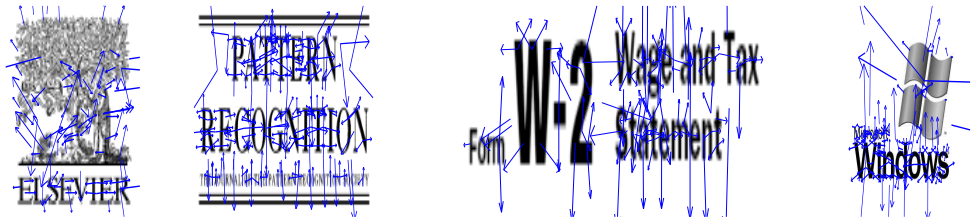


Figure 4.3 SIFT features from example logo classes.

4.3 Inverted Index Computation

In order to avoid matching query document features with features in all logo classes, inverted index [63, 61, 35] data structure is adapted to efficiently match with logo classes. This section presents an approach to store SIFT features extracted from logo classes and other related information to enhance query document class prediction. The inverted index computation is performed off-line. Let $L = \{L_1, L_2, ..., L_n\}$ be the set of logo classes to be indexed. For each logo class $L_i \in L$, $1 \leq i \leq n$, repeat the following steps.

1. Feature Extraction: Extract SIFT features [28] $X_i = \{(x_i^j, y_i^j, \{f_i^j\})\}$, $1 \leq j \leq m_i$; where m_i is the total number of features extracted from logo class L_i; (x_i^j, y_i^j) denotes feature position in L_i and $\{f_i^j\}$ denotes set of all corresponding d-dimensional descriptions as mentioned in Section 4.2.

2. Feature Quantization: Compute visual word vocabulary $C = \{C_k\}$, $1 \leq k \leq K$; where K is the size of the vocabulary, C_k is k^{th} cluster centroid; by subjecting a hundred thousand SIFT feature descriptors that arise from query document collection to K-means [64] clustering. These descriptors are not just limited to logos and represent the information from text, figures, etc. For each feature point X^j in X_i, $1 \leq j \leq m_i$; compute set of visual words $\{w_i^j\}$ by quantizing [63] set of all associated d-dimensional feature descriptors $\{f_i^j\}$ using vocabulary C. While quantizing, along with visual words $\{w_i^j\}$, compute corresponding Hamming Embedding (HE) [34] $\{he_i^j\}$ using Equation 4.1, which provides an encoding of the descriptor in the

corresponding cluster. Update X_i as $\{(x_i^j, y_i^j, \{w_i^j\}, \{he_i^j\})\}$; .

$$\begin{aligned} he_i^j(x) &= 1, \quad if\ f_i^j(x) <= C_{w_i^j}(x);\ 1 \leq x \leq d \\ &= 0, \quad otherwise; \end{aligned} \tag{4.1}$$

3. Inverted Indexing: For each visual word w_i^j in X_i, compute indexed feature as shown in Figure 4.4 and attach it to inverted index I at visual word w_i^j. **Logo ID** is logo class ID i.e., i, **Feature ID** is the SIFT feature in which the visual word w_i^j appears i.e., j, **HE** is he_i^j, and **Number of feature words** is the set cardinality $|\{w_i^j\}|$ of feature X_i^j.

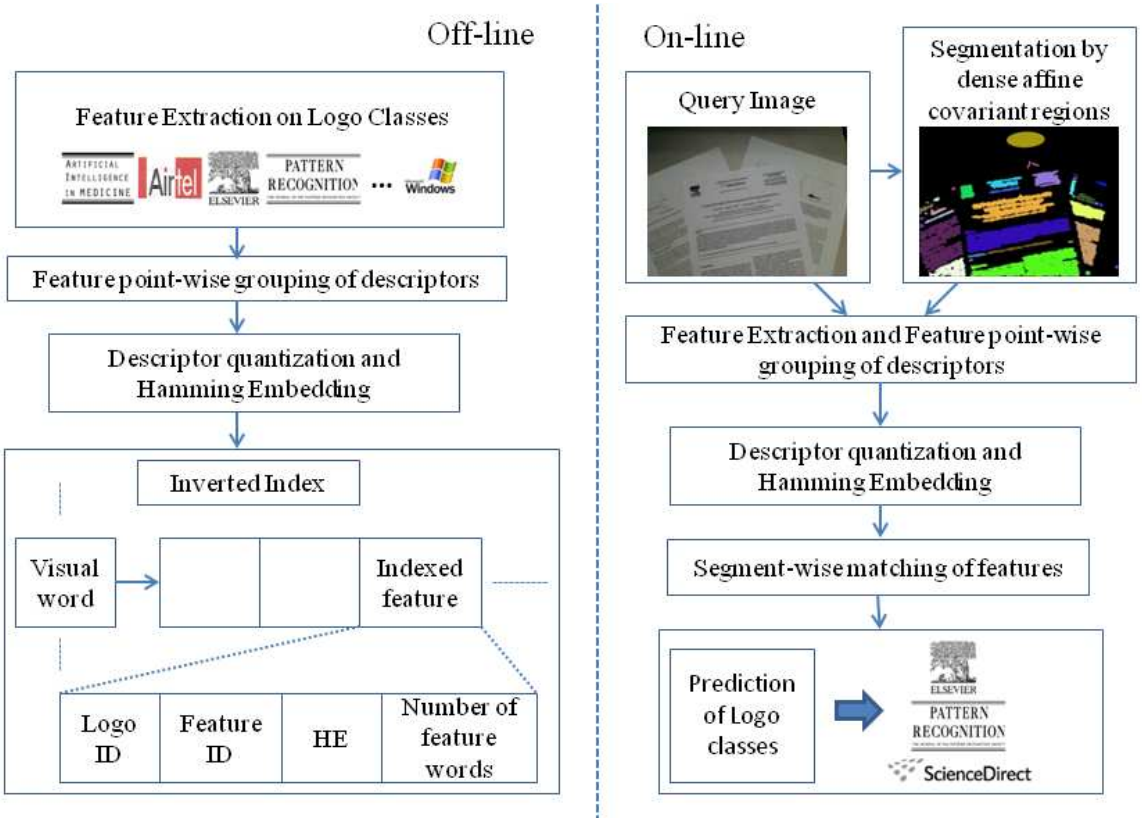


Figure 4.4 Document categorization framework.

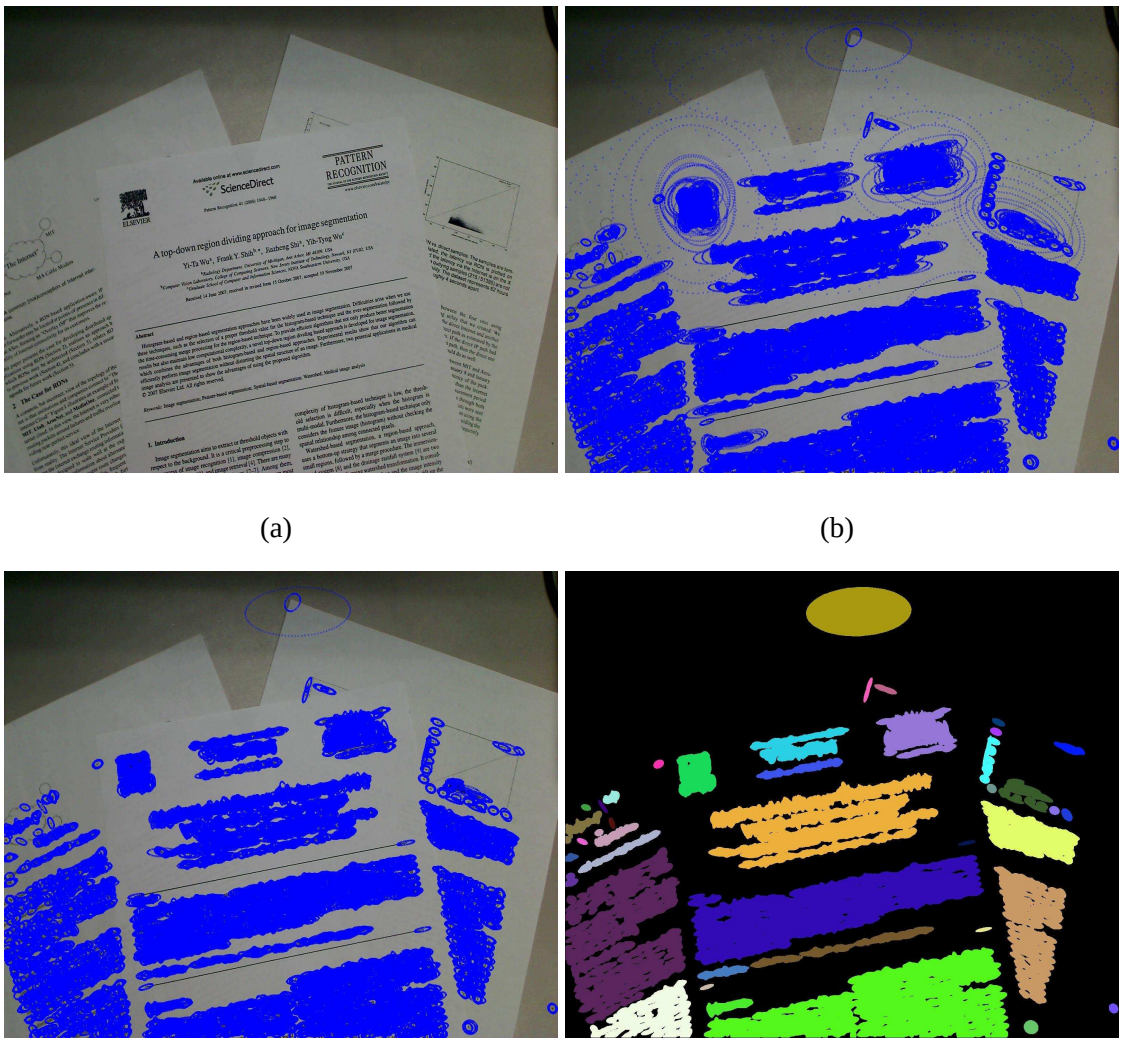
4.4 Categorization of Query Document

Generally, query document images contain a lot of data other than logos such as text, tables, figures, etc. Computing matching scores using extracted features from the entire document with the computed inverted index I could spoil logo classes prediction as shown in Figure 4.1. So, content on query document is segmented using dense affine-covariant regions, such as Hessian-Affine [30] and Harris-Affine [30]. Hessian-Affine regions are selected as they produce more repeatable regions compared to Harris-Affine regions [30]. Hessian-Affine regions are defined by affine normalization around Hessian points [30]. An iterative estimation of elliptical affine region around Hessian interest points is performed using autocorrelation matrix [45, 46]. The following subsections briefly present the methodology of segmentation and underlying logo classes prediction in the query document.

4.4.1 Segmentation

Extract Hessian-Affine regions $R = \{(x_r^j, y_r^j, a_r^j, b_r^j, c_r^j)\}$, $1 \leq j \leq m_r$ from query document image; where m_r is the total number of regions extracted, (x_r^j, y_r^j) denotes feature point position, and (a_r^j, b_r^j, c_r^j) is the corresponding region representation as ellipse. As large regions are less repeatable to view-point and illumination variations, eliminate regions in R which contain more than two other regions. Grouping the area of all ellipses that intersect with each other yields seg_q number of segments which are quite separated from each other in query document. Figure 4.5 shows a segmented image achieved by grouping dense ellipses. Figure 4.5(b) shows Hessian-Affine regions extracted from a query image shown in Figure 4.5(a). Figure 4.5(c) shows the corresponding Hessian-Affine regions remained after the elimination of regions that contain more than two regions. Finally, Figure 4.5(d) shows the segments achieved by grouping the areas of all regions that intersect with each other. Figure 4.6 shows segmentations achieved on some challenging images from the data set. First column of Figure 4.6 corresponds to original camera captured documents and the

corresponding segmented images are shown in second column. Figure 4.6 shows that the logos are quite separated from other content and entire logo area fall under same segment. Though some non logo regions are also included in the same segment corresponding to logo, the area of the segment is still much less than the entire query document and solves the purpose of using features close to the logo region for matching. Let $P = \{p^j\}$, $1 \leq j \leq seg_q$ be the set of polygons obtained by approximating each of the segment contours with a polygon [108].



(a) (b)

(c) (d)

Figure 4.5 (a) query image, (b) affine covariant regions, (c) refined regions, and (d) segmentation after grouping area under intersecting regions.

4.4.2 Feature Extraction, Quantization and Segment-wise Grouping

Extract SIFT features [28] $Q = \{(x^j, y^j, \{f^j\})\}$, $1 \leq j \leq m_q$; where m_q is the total number of features extracted from the query document; (x^j, y^j) denotes feature position and $\{f^j\}$ denotes set of all corresponding d-dimensional descriptions as mentioned in Section 4.2. Perform feature quantization on Q as presented in Section 4.3 to update Q as $\{(x^j, y^j, \{w^j\}, \{he^j\})\}$.

Divide Q into seg_q groups by assigning each query feature in Q^j to one of the segment polygons P^j that contain corresponding feature point (x^j, y^j). Denote the resulting feature groups as Q_g, $1 \leq g \leq seg_q$.

4.4.3 Matching and Score Computation

For each group of features Q_g, $1 \leq g \leq seg_q$, repeat the following steps:

1. Scores Initialization: Initialize segment scores of indexed logo classes $S_g = \{s_g^i\}$, $1 \leq i \leq n$ to zero; where n is the total number of indexed logo classes.

2. Matching: Experiments are conducted with the following two types of matching.

   (i) With out using HE: Parse inverted index I for each feature $Q_g^j \in Q_g$ and update score s_g^i, if a feature in i^{th} logo class completely intersects with query feature Q_g^j as specified in Equation 4.2. While parsing inverted index I, buffer all logo classes along with feature numbers i.e., **Feature ID** in which the corresponding visual word appears. Consider only those logo classes as a match which has a feature that exactly contains same set of visual words as query feature.

$$s_g^i \; += 1, \quad if \; \frac{|\{w_g^j\} \cap \{w_i^j\}|}{|\{w_g^j\} \cup \{w_i^j\}|} \; = \; 1 \tag{4.2}$$

   (ii) Using HE: Score update is performed similar to above matching method, except individual visual words match is refined using hamming distance as specified in

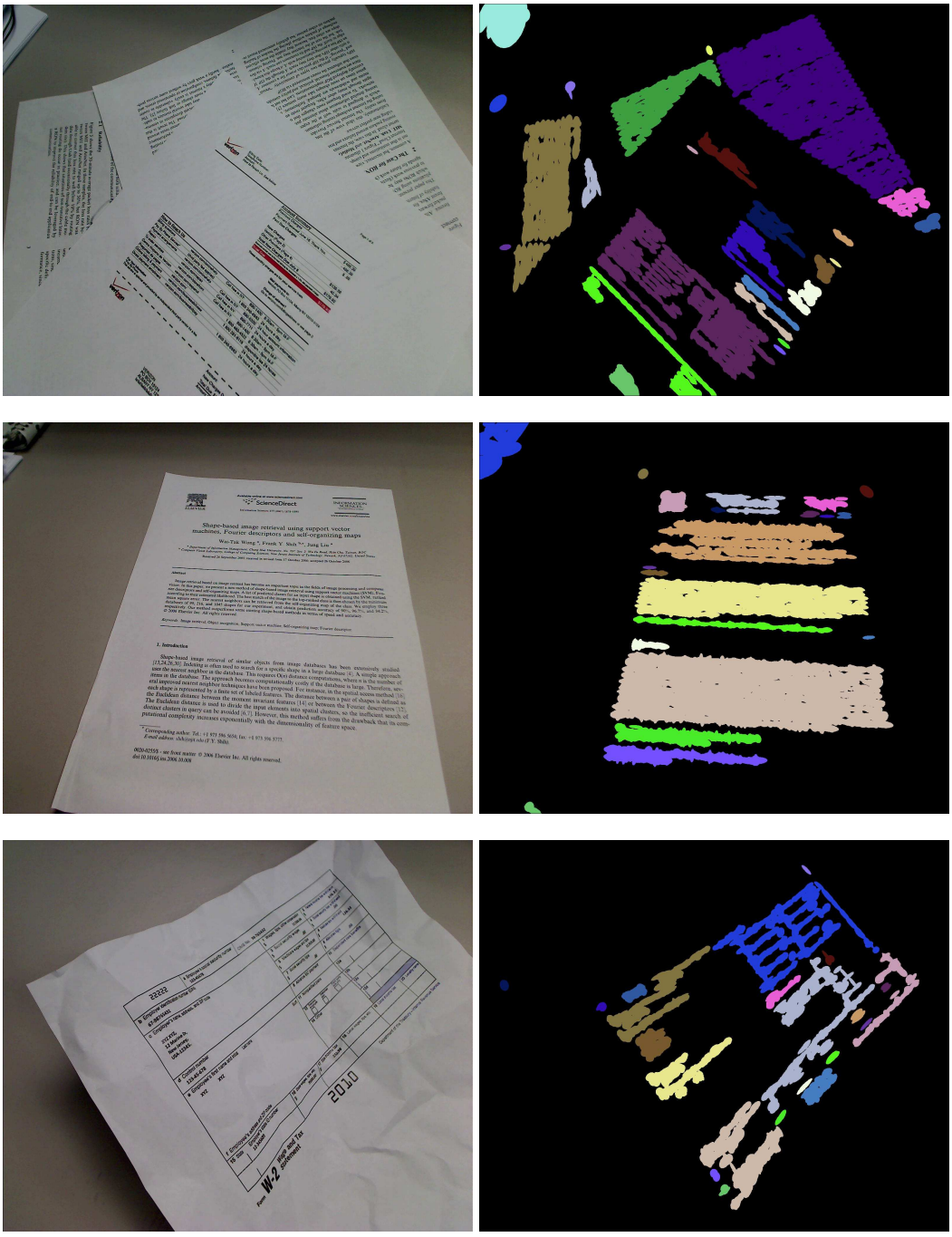


Figure 4.6 Segmentation on challenging images: original images are shown in first column and corresponding segmentation images are shown in second column.

Equation 4.3.

$$\begin{aligned} Hamming\_match(he_g^j, he_i^j) &= 1, \quad if\ xor(he_g^j, he_i^j) \leq h_t \\ &= 0, \quad otherwise \end{aligned} \tag{4.3}$$

where h_t is hamming distance threshold. h_t is set to 22 as suggested in [34] for 128-dimensional SIFT description in the experiments.

Compute document-wise logo class scores $S = \{s^i\}$, $1 \leq i \leq n$ as $s^i = max(s_g^i)$, $1 \leq g \leq seg_q$. This step updates the retrieved logo classes scores with the peak segment score associated with the corresponding logo class. Sort scores S in descending order, and categorize query document into all logo classes that have a score which is not less than t_p of the top logo class score. The impact of t_p is briefly explained in the experimental results section.

4.5 Experimental Results and Discussion

Test set consists of 300 query documents of resolution 1600 $\times$ 1200 belonging to 25 logo classes captured using Logitech Webcam Pro 9000. As some query documents also contain more than one logo class e.g., scientific articles, Figure 4.7 shows logo classes and their distribution in the test set. The test set is composed of very challenging documents such as illumination and view-point variations, cluttering, and crumples as shown in Figure 3.1. Recall [64], precision [64], and F–measure [64] are measured as defined in Equations 4.4, 4.5, and 4.6, respectively for each query document and average them over all 300 documents to produce average recall, average precision, and average F–measure accuracies. The higher these measures are, the better the prediction. Following experiments are conducted using 128 dimensional SIFT description and vocabulary sizes of 100, 500, and 1000 computed using K-means clustering of one hundred thousand SIFT descriptors extracted from the test set.

$$Recall = \frac{Number\ of\ true\ logo\ classes\ predicted}{Total\ number\ of\ true\ logo\ classes} \tag{4.4}$$

(a) 12 documents

(b) 12 documents

(c) 12 documents

(d) 12 documents (e) 108 documents (f) 36 documents (g) 84 documents

Figure 4.7 Logo classes and their distribution in query documents dataset.

$$Precision = \frac{Number\ of\ true\ logo\ classes\ predicted}{Total\ number\ of\ predicted\ logo\ classes} \quad (4.5)$$

$$F\text{–}measure = 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (4.6)$$

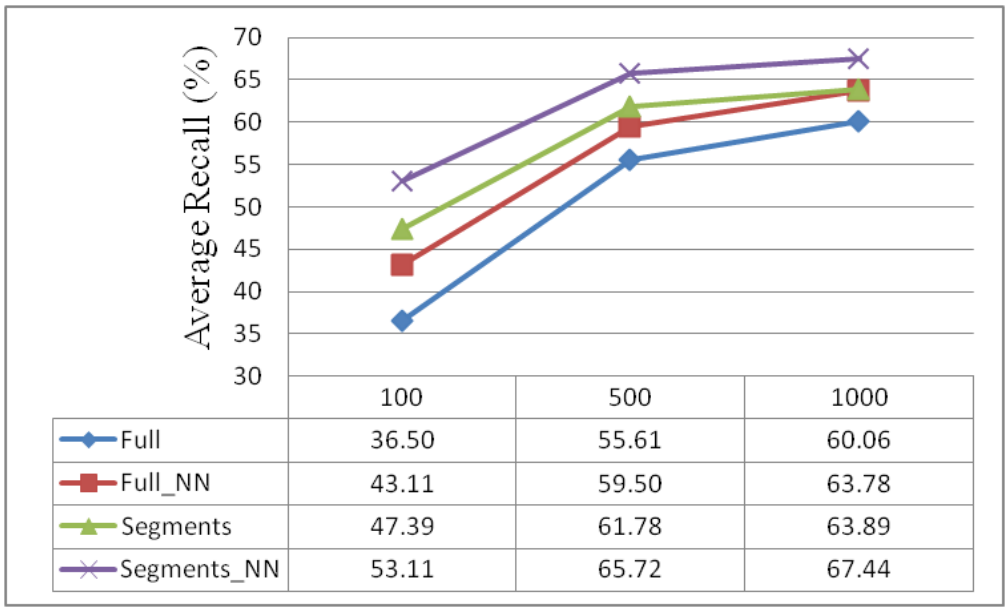
Eight matching methods are compared using SIFT features: (i) Full: Matching SIFT features from the entire query document image, (ii) Full_HE: Matching SIFT features from the entire query document image using Hamming Embedding, (iii) Full_ NN: While matching SIFT features from the entire query document image, only discriminative matches are considered i.e. if a feature has more than one match in the same logo class then the corresponding matches will be discarded, (iv) Full_NN_HE: Similar to Full_NN except that HE is enforced during matching, (v) Segments: Segment-wise matching of SIFT features from the entire query document image as presented in Section 4.4.3 with out using HE, (vi) Segments_HE: Similar to Segments with additional match refinement using HE, (vii) Segments_NN: While matching SIFT features segment-wise, only discriminative matches are considered as mentioned in Full_NN, and (viii) Segments_NN_HE: HE is applied while performing Segments_NN. Figures 4.8 and 4.9 show average recall and average precision accuracies at different vocabulary sizes using t_p = 0.8. Enforcing discriminative matches during matching i.e., all NN variants improves average recall, shown in Figure 4.8 and applying HE to establish a match significantly improves the average precision for all methodologies, shown in Figure 4.9. Figure 4.8 also shows that the application of HE does not significantly change average recall. However, enforcing discriminative matches considerably improves average precision, shown in Figure 4.9. Furthermore, segment-wise matching not only improves the average recall but also average precision. Similar pattern is observed at t_p = 0.6, which is shown in Figures 4.10 and 4.11. As more number of false predictions fall by reducing the threshold t_p to 0.6 the average precision accuracies are lower than those at t_p = 0.8, which provokes higher average recall

accuracies at t_p = 0.6. Additionally, both average recall and average precision accuracies increase as the vocabulary size increases.

Tables 4.1 and 4.2 list the F–measure accuracies at t_p = 0.8 and t_p = 0.6, respectively. The F–measure accuracies improve by 3.15% to 13.25% using segment-wise matching i.e., Segment_NN_HE compared to entire query document features i.e., Full_NN_HE. Though the F–measure accuracies are slightly better at t_p = 0.6, the improvement is slightly higher at t_p = 0.8. The improvement in F–measure accuracy goes down as the vocabulary size increases. Figure 4.12 compares the F–measure accuracies of SIFT, Speeded-Up Robust Features (SURF) [25], Hessian-Affine [30], and Harris-Affine [30] at t_p = 0.8 and t_p = 0.6. In the case of Harris-Affine, segmentation is conducted using Harris-Affine regions. Figure 4.12 demonstrates that SIFT outperforms other feature types and SURF is the poor performer amongst all and the F–measures accuracies are slightly better at t_p = 0.6. The entire methodology is implemented in C++ using OpenCV 2.3 and the experiments are conducted on an AMD quad core Linux machine with 8GB RAM. On average, it takes 500 milli seconds to categorize a single 1600×1200 camera captured document.

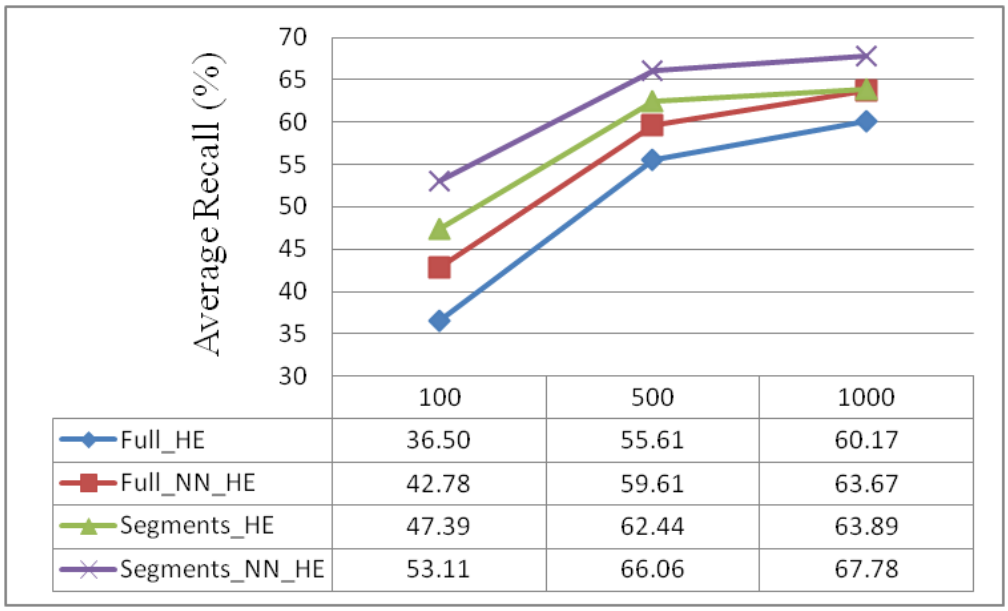
Table 4.1 F–measure Accuracies at t_p = 0.8

	Vocabulary Size		
	100	500	1000
Full	7.75%	10.29%	10.91%
Full_HE	35.03%	58.97%	64.36%
Full_NN	7.55%	7.20%	9.84%
Full_NN_HE	43.50%	63.77%	68.76%
Segments	7.09%	13.64%	16.54%
Segments_HE	49.01%	66.76%	69.60%
Segments_NN	6.57%	10.39%	13.14%
Segments_NN_HE	55.22%	71.32%	73.16%



	100	500	1000
Full	36.50	55.61	60.06
Full_NN	43.11	59.50	63.78
Segments	47.39	61.78	63.89
Segments_NN	53.11	65.72	67.44

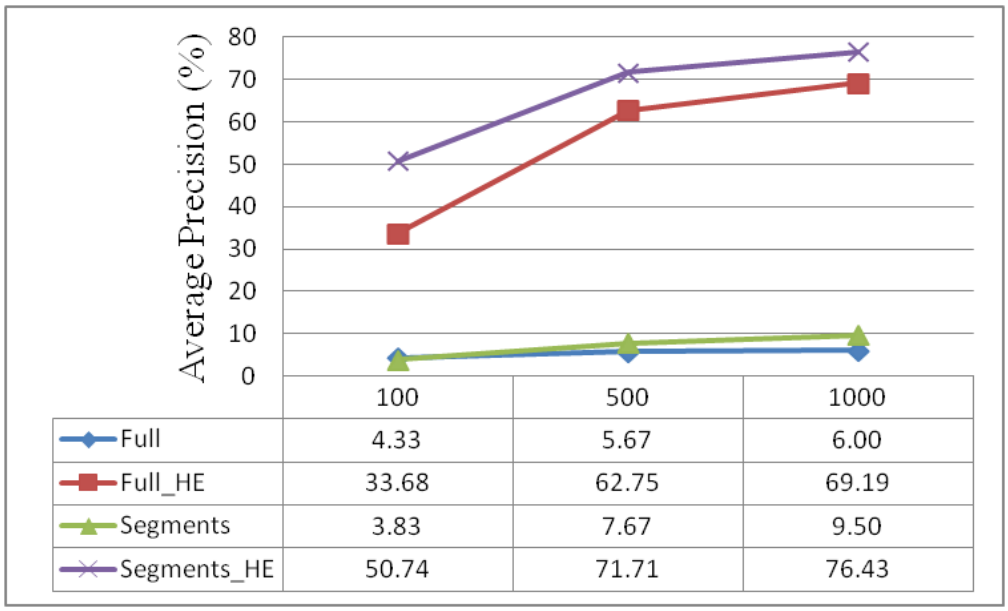
(a)



	100	500	1000
Full_HE	36.50	55.61	60.17
Full_NN_HE	42.78	59.61	63.67
Segments_HE	47.39	62.44	63.89
Segments_NN_HE	53.11	66.06	67.78

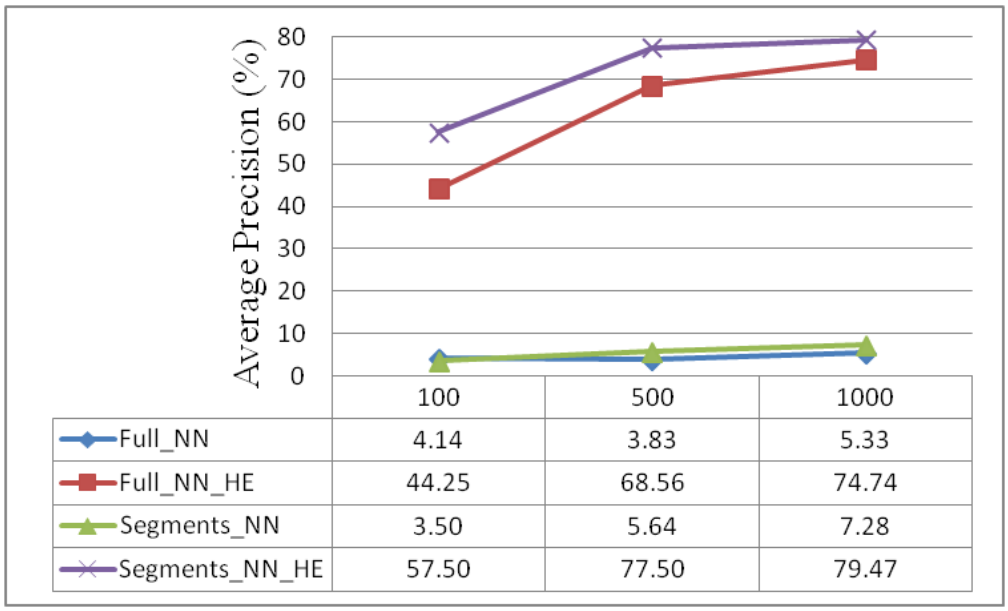
(b)

Figure 4.8 Average recall accuracies at t_p = 0.8.



	100	500	1000
Full	4.33	5.67	6.00
Full_HE	33.68	62.75	69.19
Segments	3.83	7.67	9.50
Segments_HE	50.74	71.71	76.43

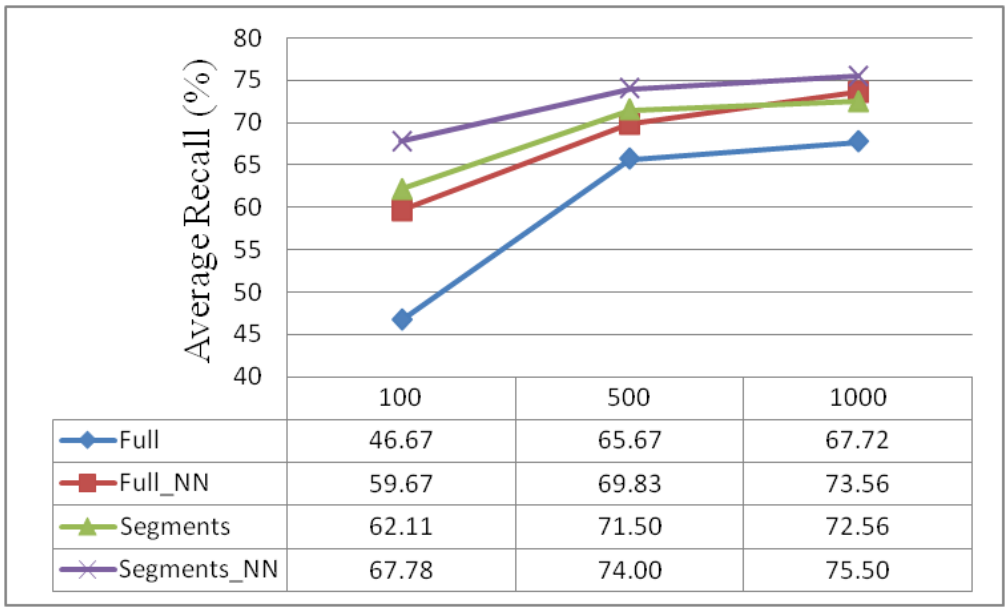
(a)



	100	500	1000
Full_NN	4.14	3.83	5.33
Full_NN_HE	44.25	68.56	74.74
Segments_NN	3.50	5.64	7.28
Segments_NN_HE	57.50	77.50	79.47

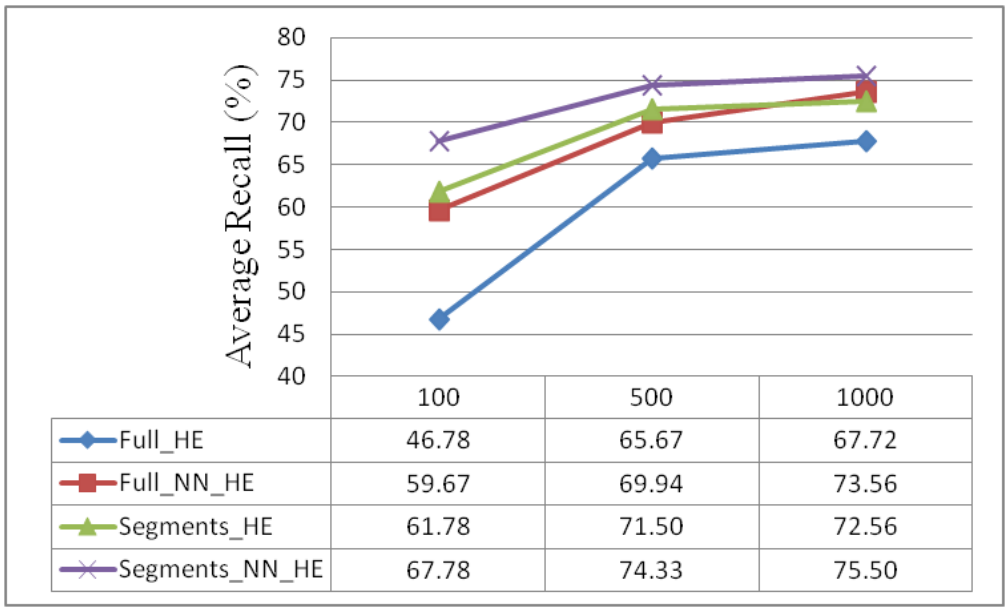
(b)

Figure 4.9 Average precision accuracies at t_p = 0.8.



	100	500	1000
Full	46.67	65.67	67.72
Full_NN	59.67	69.83	73.56
Segments	62.11	71.50	72.56
Segments_NN	67.78	74.00	75.50

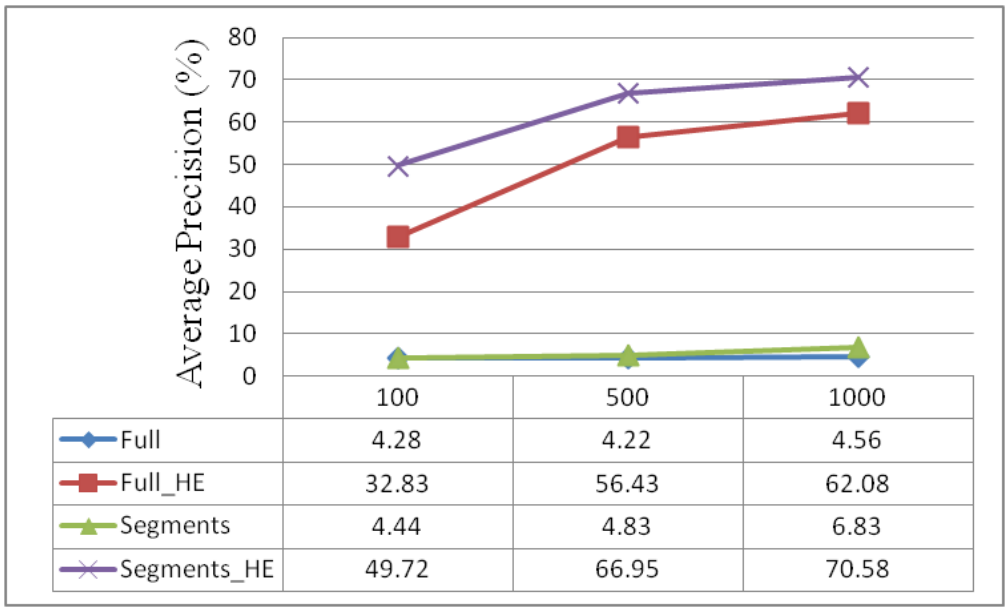
(a)



	100	500	1000
Full_HE	46.78	65.67	67.72
Full_NN_HE	59.67	69.94	73.56
Segments_HE	61.78	71.50	72.56
Segments_NN_HE	67.78	74.33	75.50

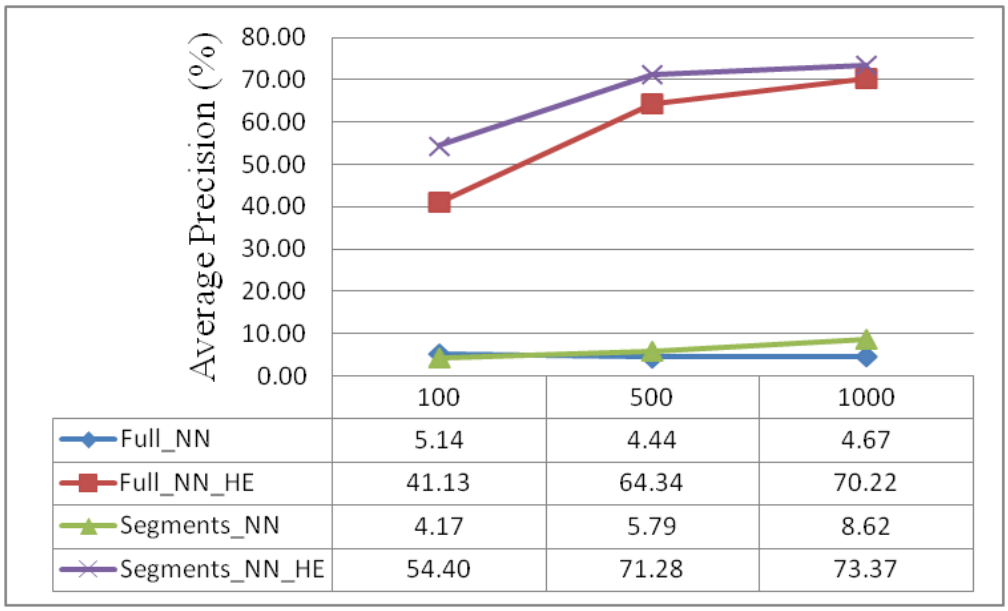
(b)

Figure 4.10 Average recall accuracies at t_p = 0.6.



	100	500	1000
Full	4.28	4.22	4.56
Full_HE	32.83	56.43	62.08
Segments	4.44	4.83	6.83
Segments_HE	49.72	66.95	70.58

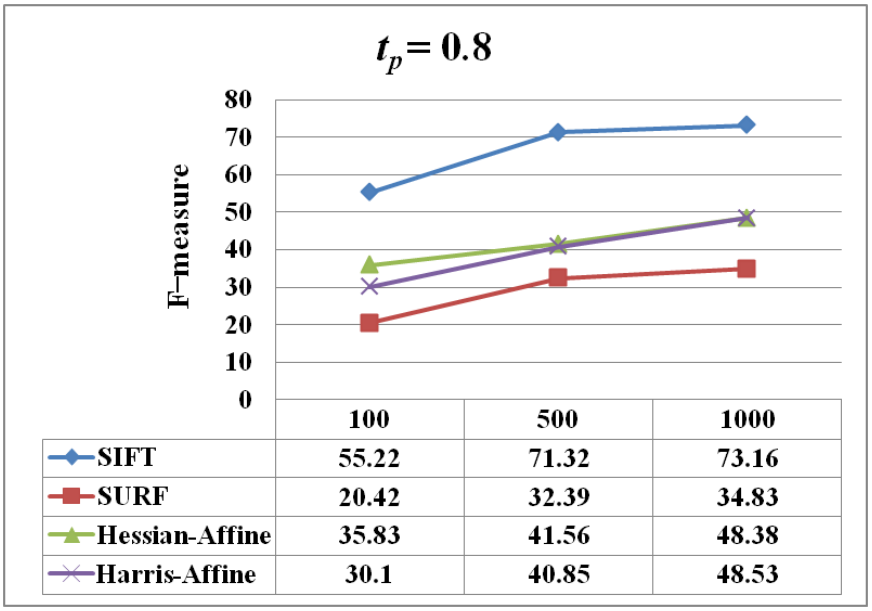
(a)



	100	500	1000
Full_NN	5.14	4.44	4.67
Full_NN_HE	41.13	64.34	70.22
Segments_NN	4.17	5.79	8.62
Segments_NN_HE	54.40	71.28	73.37

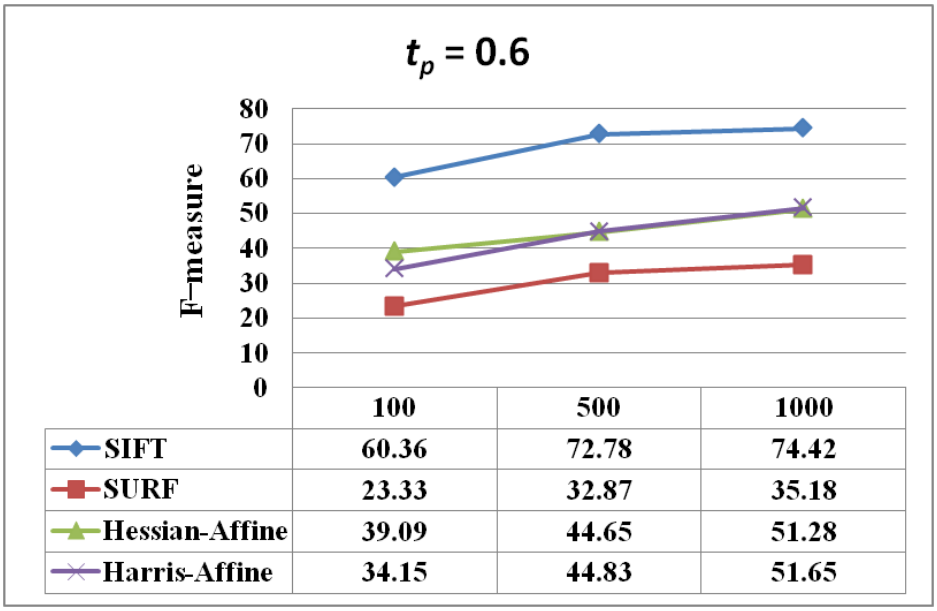
(b)

Figure 4.11 Average precision accuracies at t_p = 0.6.



	100	500	1000
SIFT	55.22	71.32	73.16
SURF	20.42	32.39	34.83
Hessian-Affine	35.83	41.56	48.38
Harris-Affine	30.1	40.85	48.53

(a)



	100	500	1000
SIFT	60.36	72.78	74.42
SURF	23.33	32.87	35.18
Hessian-Affine	39.09	44.65	51.28
Harris-Affine	34.15	44.83	51.65

(b)

Figure 4.12 F–measure accuracies of different feature types at (a) t_p = 0.8 and (b) t_p = 0.6.

Table 4.2 F–measure Accuracies at t_p = 0.6

	Vocabulary Size		
	100	500	1000
Full	7.84%	7.93%	8.54%
Full_HE	38.58%	60.70%	64.78%
Full_NN	9.46%	8.36%	8.78%
Full_NN_HE	48.70%	67.02%	71.85%
Segments	8.30%	9.05%	12.49%
Segments_HE	55.09%	69.15%	71.55%
Segments_NN	7.85%	10.74%	15.48%
Segments_NN_HE	60.36%	72.78%	74.42%

4.6 Conclusions

An affine covariant region driven segmentation approach to categorize camera captured documents by identifying logos is presented. The presented methodology not only helps to improve prediction accuracies but also gives an approximate location of the underlying logo classes in the query document, which is critical to establish correspondences for applications like registration and mosaicing. Hamming Embedding (HE) and discriminative matches are applied to increase average precision and average recall accuracies, respectively. Experimental results on a dataset of real camera captured documents demonstrated a 13.25% increase in the F–measure accuracy by computing segment-wise matching scores. Though the presented segmentation is reasonable, a more robust segmentation is desired to improve the prediction accuracies.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

In recent years, camera captured document image analysis is drawing more attention of researches due to the rapid development of inexpensive hand-held sensors, camera enabled smart phones, tablets, and so on. As there is no constrained capturing in the real world, the captured documents suffer from illumination, scale and viewpoint variations along with clutter, occlusion, and crumples. Two high level processing tasks, registration and categorization of camera captured documents using local features are presented in this dissertation.

5.1 Summary of Contributions

The following summarizes the contributions of this dissertation:

1. A novel framework to register Regions of Interest (ROI) under non-rigid deformations is developed.
    - Clustering of feature points near ROI and histogram based refinement of outliers in the correspondence set to improve convergence of traditional iterative outlier elimination mechanisms such as RANdom SAmple Consensus (RANSAC) and Thin Plate Spline-Robust Point Matching (TPS-RPM) are embedded.
    - Enhancements to RANSAC and TPS-RPM are proposed by validating the registration parameters.
    - Behavior of SIFT and SURF with respect to proposed enhancements is presented.

2. A methodology to categorize camera captured documents into predefined logo classes is presented.

    - Robust features are derived by comparisons among various local invariant features under different criteria such as feature count, repeatability, and distinctiveness.
    - Trade-off between feature representation and categorization accuracy is demonstrated.

3. A segment-wise matching methodology to categorize camera captured documents by detecting logos is presented.

    - Segmentation of query documents using dense affine covariant regions is proposed.
    - Feature-wise grouping of descriptors is presented.
    - Experimental results on a data set of real camera captured documents achieved a peak 13.25% accuracy using segment-wise matching as compared the former approach.

5.2 Limitations and Future Work

The following lists the future work that comprises the addressing of limitations as well as the extensions of the presented work:

1. One limitation in the presented registration methodology is that the matching is applied to known ROI in the template image. While this is a reasonable assumption for several document processing applications, it is not a valid assumption in general. Future work focuses on the elastic registration [65, 74, 68, 72, 75] of entire camera captured document image by fusing page segmentation [109, 110, 111], text flow analysis [105, 112, 113, 114], and geometric rectification methods [16, 5, 8] with the

approach presented in Chapter 2. Besides, improving the registration of ROI using a short video of the document captured in different perspectives [15] is also the focus of future research.

2. Though the segmentation methodology presented in Chapter 4 improved the prediction accuracies, it is not robust to the occlusions and severe camera capturing noise. Enhancing the presented segmentation approach using document layout [115, 116, 111] and document content [117, 118, 119] is also the focus of future research.

3. Finally, future work also includes the robust text detection in natural scene images and videos [2, 14].

REFERENCES

[1] S. S. Bukhari, F. Shafait, and T. M. Breuel. Coupled snakelet model for curled textline segmentation of camera-captured document images. In *Proceedings of the Tenth International Conference on Document Analysis and Recognition*, pages 61–65, Barcelona, July 2009.

[2] H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk, and B. Girod. Robust text detection in natural images with edge-enhanced maximally stable extremal regions. In *Proceedings of the IEEE International Conference on Image Processing*, Brussels, September 2011.

[3] D. Doermann, J. Liang, and H. Li. Progress in camera-based document image analysis. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition*, pages 606–616, Edinburgh, Scotland, August 2003.

[4] J. Hannuksela, P. Sangi, J. Heikkila, X. Liu, and D. Doermann. Document image mosaicing with mobile phones. In *Proceedings of the Fourteenth International Conference on Image Analysis and Processing*, pages 575–582, Modena, September 2007.

[5] J. Liang, D. DeMenthon, and D. Doermann. Flattening curved documents in images. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition*, pages 338–345, San Diego, CA, USA, June 2005.

[6] J. Liang, D. Doermann, and H. Li. Camera-based analysis of text and documents: A survey. *International Journal on Document Analysis and Recognition*, 7(2+3):83–104, July 2005.

[7] J. Liang, D. DeMenthon, and D. Doermann. Camera-based document image mosaicing. In *Proceedings of the International Conference on Pattern Recognition*, pages 476–479, Hong Kong, August 2006.

[8] J. Liang, D. DeMenthon, and D. Doermann. Geometric rectification of camera-captured document images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):591–605, April 2008.

[9] J. Liang, D. DeMenthon, and D. Doermann. Mosaicing of camera-captured document images. *Computer Vision and Image Understanding*, 113(4):572–579, April 2009.

[10] T. Nakai, K. Kise, and M. Iwamura. Camera-based document image retrieval as voting for partial signatures of projective invariants. In *Proceedings of the Eighth International Conference on Document Analysis and Recognition*, pages 379–383, August 2005.

[11] T. Nakai, K. Kise, and M. Iwamura. Use of affine invariants in locally likely arrangement hashing for camera-based document image retrieval. In *Proceedings of the Document Analysis Systems*, pages 541–552, 2006.

[12] T. Nakai, K. Kise, and M. Iwamura. Camera-based document image mosaicing using llah. In *Proceedings of the Sixteenth Conference on Document Recognition and Retrieval*, pages 1–10, San Jose, California, USA, 2009.

[13] T. Nakai, K. Kise, and M. Iwamura. Real-time retrieval for images of documents in various languages using a web camera. In *Proceedings of the Tenth International Conference on Document Analysis and Recognition*, pages 146–150, Barcelona, July 2009.

[14] J. Park, G. Lee, E. Kim, J. Lim, S. Kim, H. Yang, M. Lee, and S. Hwang. Automatic detection and recognition of korean text in outdoor signboard images. *Pattern Recognition Letters*, 31(12):1728–1739, September 2010.

[15] S. Uchida, H. Miyazaki, and H. Sakoe. Mosaicing-by-recognition for video-based text recognition. *Pattern Recognition*, 41(4):1230–1240, April 2008.

[16] B. Fu, W. Li, M. Wu, R. Li, and Z. Xu. A document rectification approach dealing with both perspective distortion and warping based on text flow curve fitting. *International Journal of Image and Graphics*, 12(1):20–44, 2012.

[17] J. Li, Z. Fan, Y. Wu, and N. Le. Document image retrieval with local feature sequences. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 346–350, July 2009.

[18] J. Moraleda. Large scalability in document image matching using text retrieval. *Pattern Recognition Letters*, 33(7):863–871, May 2012.

[19] Z. Li, M. S. Austum, and M. Neschen. Fast logo detection and recognition in document images. In *Proceedings of the Twentieth International Conference on Pattern Recognition*, pages 2716–2719, Istanbul, Turkey, August 2010.

[20] M. Rusinol and J. Llados. Logo spotting by a bag-of-words approach for document categorization. In *Proceedings of the Tenth International Conference on Document Analysis and Recognition*, pages 111–115, Barcelona, Spain, July 2009.

[21] H. Wang and Y. Chen. Logo detection in document images based on boundary extension of feature rectangles. In *Proceedings of the Tenth International Conference on Document Analysis and Recognition*, pages 1335–1339, Barcelona, Spain, July 2009.

[22] H. Wang. Document logo detection and recognition using bayesian model. In *Proceedings of the Twentieth International Conference on Pattern Recognition*, pages 1961–1964, Istanbul, Turkey, August 2010.

[23] G. Zhu and D. Doermann. Automatic document logo detection. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition*, pages 864–868, Curitiba, Brazil, September 2007.

[24] H. Bay, T. Tuytelaars, and L. V. Gool. Surf: Speeded-up robust features. In *Proceedings of the Ninth European Conference on Computer Vision*, pages 404–417, Graz, Austria, May 2006.

[25] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[26] C. Harris and M. Stephens. A combined corner and edge detector. In *Proceedings of the Alvey Vision Conference*, pages 147–151, 1988.

[27] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pages 1150–1157, Kerkyra, Greece, September 1999.

[28] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[29] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, October 2004.

[30] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1-2):43–72, November 2005.

[31] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing: Analysis and Machine Vision*. Thomson Engineering, 3 edition, 2007.

[32] T. Kobayashi and N. Otsu. Bag of hierarchical co-occurrence features for image classification. In *Proceedings of the Twentieth International Conference on Pattern Recognition*, pages 3882–3885, Istanbul, August 2010.

[33] L. Setia, A. Teynor, A. Halawani, and H. Burkhardt. Image classification using cluster co-occurrence matrices of local features. In *Proceedings of the Eighth ACM International Workshop on Multimedia Information Retrieval*, pages 173–182, Santa Barbara, California, USA, 2006.

[34] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proceedings of the Tenth European Conference on Computer Vision: Part I*, pages 304–317, Marseille, France, October 2008.

[35] Z. Wu, Q. ke, M. Isard, and J. Sun. Bundling features for large scale partial-duplicate web image search. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 25–32, Miami, FL, USA, June 2009.

[36] E. Maani, S. A. Tsaftaris, and A. K. Katsaggelos. Local feature extraction for video copy detection in a database. In *Proceedings of the Fifteenth IEEE International Conference on Image Processing*, pages 1716–1719, San Diego, CA, October 2008.

[37] J. L. To, L. Chen, A. Joly, I. Laptev, O. Buisson, V. G. Brunet, N. Boujemaa, and F. Stentiford. Video copy detection: A comparative study. In *Proceedings of the Sixth International Conference on Image and Video Retrieval*, pages 371–378, Amsterdam, The Netherlands, 2007.

[38] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10):761–767, June 2004.

[39] Q. Fan, K. Barnard, A. Amir, and A. Efrat. Robust spatiotemporal matching of electronic slides to presentation videos. *IEEE Transactions on Image Processing*, 20(8):2315–2328, 2011.

[40] B. Klare, Z. Li, and A. K. Jain. Matching forensic sketches to mug shot photos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):639–646, 2011.

[41] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):978–994, 2011.

[42] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.

[43] S. Ehsan, N. Kanwal, A. F. Clark, and K. D. McDonald-Maier. An algorithm for the contextual adaptation of surf octave selection with good matching performance:best octaves. *IEEE Transactions on Image Processing*, 21(1):297–304, 2012.

[44] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1265–1278, 2005.

[45] A. Baumberg. Reliable feature matching across widely separated views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 774–781, Hilton Head Island, South Carolina, USA, 2000.

[46] T. Lindeberg and J. Garding. Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure. *Image and Vision Computing*, 15(6):415–434, 1997.

[47] R. Kimmel, C. Zhang, A. M. Bronstein, and M. M. Bronstein. Are mser features really interesting? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2316–2320, 2011.

[48] B. Jian and B. C. Vemuri. Robust point set registration using gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1633–1645, 2011.

[49] S. Liao and A. C. S. Chung. Nonrigid brain mr image registration using uniform spherical region descriptor. *IEEE Transactions on Image Processing*, 21(1):157–169, 2012.

[50] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, November 2000.

[51] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2 edition, 2003.

[52] S. Choi, T. Kim, and W. Yu. Performance evaluation of ransac family. In *Proceedings of the British Machine Vision Conference*, pages 1–12, London, September 2009.

[53] Y. Li, L. Gu, and T. Kanade. Robustly aligning a shape model and its application to car alignment of unknown pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(9):1860–1876, 2011.

[54] O. Naroditsky, X. S. Zhou, J. Gallier, and S. I. Roumeliotis. Two efficient solutions for visual odometry using directional correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):818–824, 2012.

[55] H. Chui and A. Rangarajan. A feature registration framework using mixture models. In *Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, pages 190–197, Hilton Head Island, South Carolina, June 2000.

[56] H. Chui and A. Rangarajan. A new point matching algorithm for non-rigid registration. *Computer Vision and Image Understanding*, 89(2-3):114–141, February 2003.

[57] R. Dahyot. Statistical hough transform. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(8):1502–1509, 2009.

[58] J. Gall, A. Yao, N. Razavi, L. V. Gool, and V. Lempitsky. Hough forests for object detection, tracking, and action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2188–2202, 2011.

[59] F. L. Bookstein. Principle warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, June 1989.

[60] H. Jegou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2011.

[61] I. H. Witten, A. Moffat, and T. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishers, 1999.

[62] Z. Wu, Q. Ke, J. Sun, and H. Shum. Scalable face image retrieval with identity-based quantization and multireference reranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1991–2001, 2011.

[63] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings of the Ninth International Conference on Computer Vision*, pages 1470–1477, Nice, France, October 2003.

[64] D. L. Olsen and D. Delen. *Advanced Data Mining Techniques*. Springer, Edition 1, Miami, FL, USA, February 2008.

[65] C. Domokos, J. Nemeth, and Z. Kato. Nonlinear shape registration without correspondences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):943–958, 2012.

[66] Y. Wang, K. Liu, Q. Hao, X. Wang, D. L. Lau, and L. G. Hassebrook. Robust active stereo vision using kullback-leibler divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):548–563, 2012.

[67] R. Horaud, F. Forbes, M. Yguel, G. Dewaele, and J. Zhang. Rigid and articulated point registration with expectation conditional maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):587–602, 2011.

[68] Y. Liu. Penalizing closest point sharing for automatic free form shape registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):1058–1064, 2011.

[69] J. Lee and C. Won. Topology preserving relaxation labeling for nonrigid point matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2):427–432, 2011.

[70] C. Xing and P. Qiu. Intensity-based image registration by nonparametric local smoothing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):2081–2092, 2011.

[71] S. M. S. Nejhum, Y. Chi, J. Ho, and M. Yang. Higher-dimensional affine registration and vision applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7):1324–1338, 2011.

[72] D. Mahapatra and Y. Sun. Integrating segmentation information for improved mrf-based elastic image registration. *IEEE Transactions on Image Processing*, 21(1):170–183, 2012.

[73] M. M. Fouad, R. M. Dansereau, and A. D. Whitehead. Image registration under illumination variations using region-based confidence weighted m-estimators. *IEEE Transactions on Image Processing*, 21(3):1046–1060, 2012.

[74] M. Kim, G. Wu, P. Yap, and D. Shen. A general fast registration framework by learning deformation-appearance correlation. *IEEE Transactions on Image Processing*, 21(4):1823–1833, 2012.

[75] F. Zhou, W. Yang, and Q. Liao. A coarse-to-fine subpixel registration method to recover local perspective deformation in the application of image super-resolution. *IEEE Transactions on Image Processing*, 21(1):53–66, 2012.

[76] A. Mahmood and S. Khan. Correlation-coefficient-based fast template matching through partial elimination. *IEEE Transactions on Image Processing*, 21(4):2099–2108, 2012.

[77] D. Zosso, X. Bresson, and J. Thiran. Geodesic active fields-a geometric framework for image registration. *IEEE Transactions on Image Processing*, 20(5):1300–1312, 2011.

[78] I. Bazzi, R. Schwartz, and J. Makhoul. An omnifont open-vocabulary ocr system for english and arabic. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(6):495–504, 1999.

[79] Y. Xu and G. Nagy. Prototype extraction and adaptive ocr. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1280–1296, 1999.

[80] J. Park, V. Govindaraju, and S. N. Srihari. Ocr in a hierarchical feature space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(4):400–407, 2000.

[81] J. Kleban, X. Xie, and W. Ma. Spatial pyramid mining for logo detection in natural scenes. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1077–1080, Hannover, Germany, June 2008.

[82] R. Phan and D. Androutsos. Content-based retrieval of logo and trademarks in unconstrained color image databases using color edge gradient co-occurrence histograms. *Pattern Recognition*, 114(1):66–84, January 2010.

[83] S. Lu, L. Li, and C. L. Tan. Document image retrieval through word shape coding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1913–1918, 2008.

[84] G. Zhu, Y. Zheng, D. Doermann, and S. Jaeger. Signature detection and matching for document image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2015–2031, 2009.

[85] M. Jain, H. Jegou, and P. Gros. Asymmetric hamming embedding: Taking the best of our bits for large scale image search. In *Proceedings of the Nineteenth ACM International Conference on Multimedia*, pages 1441–1444, Scottsdale, Arizona, USA, November 2011.

[86] F. Y. Shih and S. Chen. Adaptive document block segmentation and classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 26(5):797–802, 1996.

[87] O. Altamura, F. Esposito, and D. Malerba. Wisdom++ : An interactive and adaptive document analysis system. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 366–369, 1999.

[88] M. Diligenti, P. Frasconi, and M. Gori. Hidden tree markov models for document image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4):519–523, 2003.

[89] X. Li and S. Lei. Block-based segmentation and adaptive coding for visually lossless compression of scanned documents. In *Proceedings of the International Conference on Image Processing*, pages 450–453, 2001.

[90] Y. Wang, I. T. Philips, and R. M. Haralick. Document zone content classification and its performance evaluation. *Pattern Recognition*, 39(1):57–73, 2006.

[91] F. Isgro and M. Pilu. A fast and robust image registration method on an early consensus paradigm. *Pattern Recognition Letters*, 25(8):943–954, 2004.

[92] A. Myronenko and X. Song. Point set registration: Coherent point drift. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 32(12):2262–2275, December 2010.

[93] M. Sofka, Y. Gehua, and C. V. Stewart. Simultaneous covariance driven correspondence (cdc) and transformation estimation in the expectation maximization framework. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–8, Minneapolis, MN, June 2007.

[94] J. Yang, R. S. Blum, J. P. Williams, Y. Sun, and C. Xu. Non-rigid image registration using geometric features and local salient region features. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 825–832, New York, NY, June 2006.

[95] M. Calonder, V. Lepetit, and P. Fua. Keypoint signatures for fast learning and recognition. In *Proceedings of the Tenth European Conference on Computer Vision*, pages 58–71, Marseille, France, October 2008.

[96] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, October 2005.

[97] J. Cha. Seamless and fast panoramic image stitching. In *Proceedings of the IEEE International Conference on Consumer Electronics*, pages 29–30, Las Vegas, January 2012.

[98] H. I. Koo and N. I. Cho. Feature-based image registration algorithm for image stitching applications on mobile devices. *IEEE Transactions on Consumer Electronics*, 57(5):1303–1310, 2011.

[99] A. Bishnu and B. B. Bhattacharya. Stacked euler vector (serve): A gray-tone image feature based on bit-plane augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2):350–355, 2007.

[100] R. G. D. Cruz, E. A. Albacea, and V. Y. Mariano. A coarse-to-fine document image registration for an automated form reader. In *Proceedings of the Eighth National Conference on IT Education*, Boracay Island, October 2010.

[101] R. Safari, N. Narasimhamurthi, M. Shridhar, and M. Ahmadi. Form registration: A computer vision approach. In *Proceedings of the Fourth International Conference on Document Analysis and Recognition*, pages 758–761, Ulm, Germany, August 1997.

[102] S. L. Taylor and R. Fritzson. Registration and region extraction of data from forms. In *Proceedings of the Eleventh International Conference on Pattern Recognition*, pages 173–176, The Hague, Netherlands, August 1992.

[103] Y. Zhu, R. Dai, B. Xiao, and C. Wang. Document image registration based on geometric invariant and contour matching. In *Proceedings of the International Conference on Computational Intelligence and Multimedia Applications*, pages 472–476, Sivakasi, Tamil Nadu, December 2007.

[104] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu. The analysis of a simple k-means clustering algorithm. In *Proceedings of the Sixteenth Annual Symposium on Computational Geometry*, pages 100–109, Hong Kong University of Science and Technology, June 2000.

[105] B. Bataineh, S. N. H. S. Abdullah, and K. Omar. An adaptive local binarization method for document images based on a novel thresholding method and dynamic windows. *Pattern Recognition Letters*, 32(14):1805–1813, October 2011.

[106] M. A. R. Ortegon, E. A. D. Guzman, R. Rojas, and E. Cuevas. Unsupervised measures for parameter selection of binarization algorithms. *Pattern Recognition*, 44(3):491–502, March 2011.

[107] G. Agam, S. Argamon, O. Frieder, D. Grossman, and D. Lewis. *The IIT Complex Document Image Processing(CDIP) Test Collection Project*. Illinois Institute of Technology, USA, 2006.

[108] D. Douglas and T. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *The Canadian Cartographer*, 10(2):112–122, 1973.

[109] F. Shafait, D. Keysers, and T. M. Breuel. Performance evaluation and benchmarking of six-page segmentation algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6):941–954, 2008.

[110] F. Shafait and T. M. Breuel. The effect of border noise on the performance of projection-based page segmentation methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):846–851, 2011.

[111] N. Stamatopoulos, B. Gatos, and A. Kesidis. Automatics borders detection of camera captured document images. In *Proceedings of the Second International Workshop on Camera-Based Document Analysis and Recognition*, pages 71–78, Curitiba, Brazil, 2007.

[112] H. Bunke and K. Riesen. Recent advances in graph-based pattern recognition with applications in document analysis. *Pattern Recognition*, 44(5):1057–1067, May 2011.

[113] Z. Liu, H. Zhou, and N. Yang. Semi-supervised learning for text-line segmentation. *Pattern Recognition Letters*, 31(11):1260–1273, August 2010.

[114] X. Peng, S. Setlur, V. Govindaraju, and S. Ramachandrula. Using a boosted tree classifier for text segmentation in hand-annotated documents. *Pattern Recognition Letters*, 33(7):943–950, May 2012.

[115] T. Kanungo and S. Mao. Stochastic language models for style-directed layout analysis of document images. *IEEE Transactions on Image Processing*, 12(5):583–596, 2003.

[116] S. Kumar, R. Gupta, N. Khanna, S. Chaudhury, and S. D. Joshi. Text extraction and document image segmentation using matched wavelets and mrf model. *IEEE Transactions on Image Processing*, 16(8):2117–2128, 2007.

[117] P. Chiu, F. Chen, and L. Denoue. Picture detection in document page images. In *Proceedings of the 10th ACM Symposium on Document Engineering*, pages 211–214, Manchester, United Kingdom, September 2010.

[118] S. Mandal, S. Chowdhury, A. Das, and B. Chanda. A simple and effective table detection system from document images. *International Journal on Document Analysis and Recognition*, 8(2):172–182, 2006.

[119] F. Shafait and R. Smith. Table detection in heterogeneous documents. In *Proceedings of the Ninth IAPR International Workshop on Document Analysis Systems*, pages 65–72, Boston, Massachusetts, June 2010.